UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

E.A.P. DE INGENIERÍA DE SISTEMAS

"DETECCION DE FRAUDES USANDO TECNICAS DE CLUSTERING"

TESIS

Tesis para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Mónica Tahiz Rantes García

Lizbeth María Cruz Quispe

ASESOR

Erick Vicente De Tomas

Lima – Perú

2010

Mónica Tahiz Rantes García
y
Lizbeth María Cruz Quispe

Detección de Fraudes usando Técnicas de Clustering

Tesis presentada a la Facultad de Ingeniería de Sistemas e Informática de la UNMSM, como parte de los requisitos para obtener el título de Ingeniero de Sistemas

Asesor: Mg. Erick Vicente De Tomas

Universidad Nacional Mayor de San Marcos
Facultad de Ingeniería de Sistemas e Informática
E.A.P. de Ingeniería de Sistemas
Lima - Perú
febrero - 2010

Para mi familia, en especial a mis padres, por su comprensión y apoyo brindado en todo momento.

Mónica

Para mis padres con mucho cariño por todo el apoyo que siempre me han dado.

Liz

Agradecimientos

Agradecemos a nuestros padres quienes nos han brindado su apoyo incondicional en todo momento a lo largo de nuestra carrera profesional y a nuestro asesor por su apoyo en la elaboración de esta tesina.

Resumen

El fraude con tarjetas de crédito es uno de los problemas más importantes a los que se enfrentan actualmente las entidades financieras. Si bien la tecnología ha permitido aumentar la seguridad en las tarjetas de crédito con el uso de claves PIN, la introducción de chips en las tarjetas, el uso de claves adicionales como tokens y mejoras en la reglamentación de su uso, también es una necesidad para las entidades bancarias, actuar de manera preventiva frente a este crimen. Para actuar de manera preventiva es necesario monitorear en tiempo real las operaciones que se realizan y tener la capacidad de reaccionar oportunamente frente a alguna operación dudosa que se realice.

La técnica de Clustering frente a esta problemática es un método muy utilizado puesto que permite la agrupación de datos lo que permitirá clasificarlos por su similitud de acuerdo a alguna métrica, esta medida de similaridad está basada en los atributos que describen a los objetos. Además ésta técnica es muy sensible a la herramienta Outlier que se caracteriza por el impacto que causa sobre el estadístico cuando va a analizar los datos.

Abstract

The credit card fraud is one of the most important problems currently facing financial institutions. While technology has enhanced security in credit cards with the use of PINs, the introduction of chips on the cards, the use of additional keys as tokens and improvements in the regulation of their use, is also a need for banks, act preemptively against this crime. To act proactively need real-time monitoring operations are carried out and have the ability to react promptly against any questionable transaction that takes place.

Clustering technique tackle this problem is a common method since it allows the grouping of data allowing classifying them by their similarity according to some metric, this measure of similarity is based on the attributes that describe the objects. Moreover, this technique is very sensitive to Outlier tool that is characterized by the impact they cause on the statistic when going to analyze the data.

Índice general

Índice de figuras

Índice de algoritmos

Capítulo 1

Introducción

1.1. Fraudes en el uso de tarjetas de crédito

El fraude con tarjetas de crédito es uno de los problemas más importantes a los que se enfrentan actualmente las entidades financieras. Algunas publicaciones calculan un aproximado de $2,5 mil millones de pérdidas mundiales en el año 2002 [Bhatla, 2003] y para el año 2007 se estimó que en EEUU se ha perdido $3,5 mil millones de dólares. Otros reportes, como el de Visa International, muestra que por cada dólar consumido con una tarjeta de crédito, 0.05 centavos corresponden a un fraude [Newman, 2003].

Si bien la tecnología ha permitido aumentar la seguridad en las tarjetas de crédito con el uso de claves PIN, la introducción de chips en las tarjetas, y el uso de dispositivos adicionales como tokens y tarjetas coordenadas, también es una necesidad para las entidades bancarias, actuar de manera preventiva frente a este crimen. Sobre todo si tenemos en cuenta que en algunos casos los criminales cometen este tipo de delitos sin su presencia física, sino valiéndose de los canales de atención electrónicos que ofrecen los bancos.

Existen básicamente dos formas de actuar de las personas que cometen este tipo de delitos [Garcia, 2003]: por un lado la obtención de la tarjeta física como tal y por el otro la grabación de los datos de la banda magnética para su posterior utilización, ya sea a través de una nueva tarjeta o utilizando los datos en compras realizadas a través de Internet.

En el primero de los casos, en los que los delincuentes obtienen la tarjeta física, una forma de obtenerla discretamente para así cometer su delito es la siguiente:

- En la ranura, donde se debe introducir la tarjeta, se coloca una nueva ranura que llevará una tope para que la tarjeta, al ser introducida, no llegue al cajero. De este modo se ha conseguido que la tarjeta quede atrapada, tal como se muestra en la 1.1.
- Aprovechando el hecho, uno de los delincuentes se acercará al usuario de la tarjeta y le comentará que a él le ha sucedido lo mismo, y que debe marcar una serie de números y para terminar su clave personal, que el delincuente estará mirando y memorizando.
- El siguiente paso, una vez que el dueño de la tarjeta se ha ido, confiado de que resolverán su problema, consiste en que un segundo delincuente (cómplice del primero) se acerque al cajero y retire la tarjeta, con lo que ya disponen de la tarjeta y de la clave personal.

Figura 1.1: Alteración de la ranura de un cajero automático

Otras de las formas frecuentes de actuar consisten en obtener los datos de la tarjeta y posteriormente grabarlos en otra para poder operar con ella. Existen en el mercado multitud de lectoras/grabadoras de bandas magnéticas, que facilitan a los delincuentes esta tarea.

En todos los casos no es necesario hacer una copia material o física de la tarjeta de crédito para llevar a cabo un uso fraudulento de la misma, se pueden hacer compras a través de Internet, es decir, mediante comercio electrónico utilizando el número de

tarjeta y la fecha de caducidad. Esta forma de comprar con la tarjeta de crédito es una de las más difundidas, si tomamos como ejemplo a países como España, podemos ver que más del 39 % utiliza este medio de pago [ONTSIII, 2007] para sus compras por internet tal como se muestra en la figura 1.2.

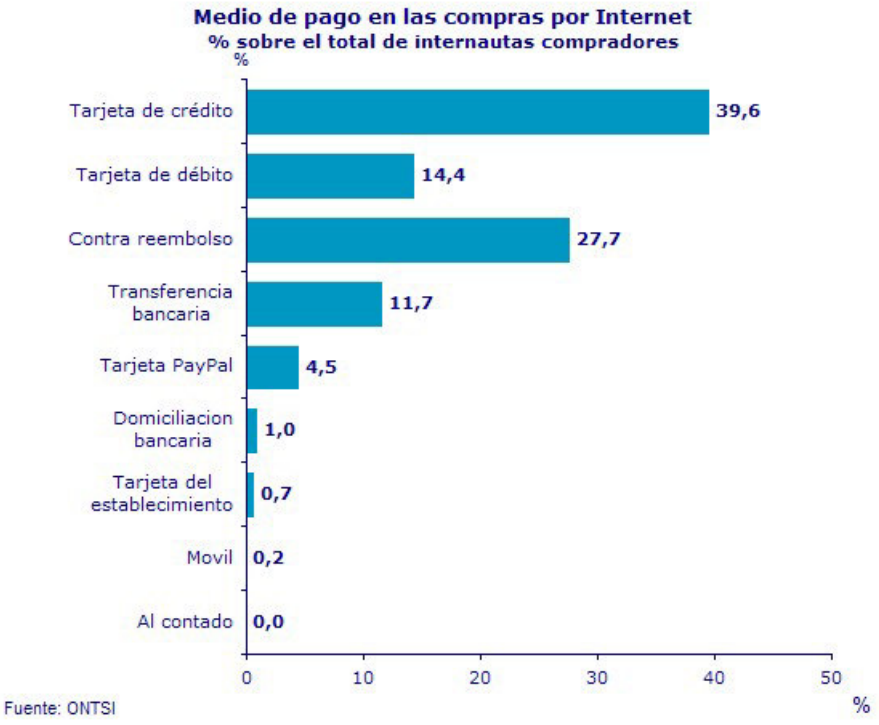

Figura 1.2: Uso de tarjeta de crédito para compras por internet en España

Para actuar de manera preventiva es necesario monitorear en tiempo real las operaciones que se realizan y tener la capacidad de reaccionar oportunamente frente a alguna operación dudosa que se realice. Esta falta de monitoreo transaccional de los canales de operación, conlleva a cierta vulnerabilidad al fraude bancario, pues a pesar que estos canales controlan el número de operaciones y el importe de las mismas, no garantizan que la persona que está realizando la operación sea un defraudador.

Las entidades financieras ante la existencia de fraudes electrónicos, solo envían comunicados de alerta a los clientes sobre las modalidades de fraude más utilizadas y recomiendan que acciones tomar para no caer en alguna de ellas. Sin embargo las entidades financieras podrían proteger a sus clientes utilizando técnicas de minería de datos y en forma automática generar alertas sobre operaciones fraudulentas. Para detectar este tipo de operaciones, primero es necesario identificar el comportamiento común de los clientes y luego detectar aquellas operaciones que salgan fuera de lo

común como posibles intentos de fraudes. Las técnicas de agrupamiento y detección de comportamiento fuera de serie que nos permiten realizar estas tareas. Estas técnicas serán revisadas en la presente disertación.

1.2. Objetivos

- Revisar un conjunto de técnicas de agrupamiento, también conocidas como clustering, que permita encontrar grupos usuarios de tarjeta de crédito sin un conocimiento previo de los mismos y a partir de datos de clientes y sus consumos, que son guardados por las entidades financieras emisoras de tarjetas de crédito.
- Revisar un conjunto de técnicas para la detección de comportamiento fuera de serie, también conocidas como outliers detection, de tal manera que permita encontrar aquellos consumos de clientes que se escapan a un patrón de comportamiento común.
- Implementar una aplicación que permita validar el método propuesto en cuanto a búsqueda de grupos y detección de comportamientos fuera de serie

1.3. Justificación

Actualmente las entidades financieras requieren explotar la información almacenada en sus bases de datos, producto de sus transacciones y operaciones diarias, con la finalidad de conocer el comportamiento del cliente y así obtener un alto nivel de competitividad y posibilidades de desarrollo, considerando para ello la prevención del fraude. Es en este contexto que Data Mining sobresale como un conjunto de técnicas para transformar y explotar los datos con el fin de extraer conocimiento útil para la toma de decisiones.

Para este tipo de necesidad han surgido soluciones de detección de comportamiento, considerando para ello las técnicas de Data Mining, utilizadas para clasificar a los clientes bajo ciertos patrones de comportamiento.

Las técnicas consideradas en el desarrollo de éste trabajo están basadas en Clustering y Detección de Outliers, ambas se complementan para poder identificar patrones de comportamiento y detectar anomalías en los mismos, los cuales son perjudiciales para el cliente y la empresa financiera, considerando para el caso de estudio el fraude por uso de tarjeta de crédito, registrándose operaciones en diferentes horarios y lugares, no realizadas por el propietario de la tarjeta.

Cabe resaltar que las técnicas en estudio permiten clasificar a los clientes, sin conocimiento previo de los mismos, lo cual facilita la toma de decisiones a la entidad financiera afectada.

Capítulo 2

Marco Teórico

2.1. Tarjetas de Crédito

2.1.1. Antecedentes

El origen de las tarjetas de crédito se remonta a las tarjetas comerciales emitidas a mediados del siglo XX por compañías de hostelería en los Estados Unidos, con el fin de facilitar a sus clientes el pago aplazado. En las décadas siguientes los grandes sistemas de tarjetas se extendieron a Europa. Y con el tiempo las entidades financieras se fueron convirtiendo en los principales emisores de las tarjetas. Surgen así las tarjetas de crédito bancarias, en las que entre el comerciante y el cliente aparece un tercero, la entidad financiera emisora de la tarjeta [Zunzunegui01].

2.1.2. Definición

La Tarjeta de Crédito es un instrumento de crédito que permite diferir el cumplimiento de las obligaciones monetarias asumidas con su sola presentación, sin la necesidad de previamente provisionar fondos a la entidad que asume la deuda, que generalmente son Bancos u otra empresa del Sistema Financiero [Patroni, 2003].

La tarjeta de Crédito es el Medio de Pago más usado, esto se debe básicamente a su fácil uso, característica esencial de este medio de pago, y por la seguridad que brinda tanto al vendedor, ya que existe alguna entidad financiera que respalda al con-

sumidor, así como para el consumidor ya que frecuentemente las Tarjetas de Crédito se encuentran amparadas por seguros. Asimismo, existe la confianza generalizada que las operaciones que se realizan utilizando Tarjetas de Crédito, están más que probadas y cuentan con todas las garantías.

En el Perú, la Tarjeta de Crédito se encuentra regulada mediante Resolución SBS N° 271-2000 - Reglamento de Tarjetas de Crédito - el cual conceptualiza la Tarjeta de Crédito como un contrato mediante el cual una empresa concede una línea de crédito al titular por un lapso determinado y entrega por tanto una tarjeta de crédito, con la finalidad que el usuario de la tarjeta adquiera bienes o servicios en los establecimientos afiliados.

En conclusión debemos decir que, la tarjeta de crédito se trata de una línea de crédito abierta a favor del cliente por una entidad emisora, esta puede ser entidades financiera supervisadas por la Superintendencia de Banca y Seguros o empresas comerciales que emiten sus propias cartas de crédito.

Es fundamental tener en cuenta que para que la Tarjeta de Crédito tenga validez, esta debe contener [Patroni, 2003]:

- La denominación de la empresa que emite la tarjeta
- El sistema de tarjeta de crédito al que pertenece
- Numeración codificada de la tarjeta
- Nombre del usuario de la tarjeta y su firma
- Fecha de vencimiento
- La indicación expresa del ámbito geográfico de validez. En caso de faltar este requisito, se entiende sin admitir prueba en contra que su validez es internacional.

2.1.3. Flujo Transactional

Este flujo o sistema se inicia cuando un consumidor obtiene una de tarjeta de crédito a través de un banco Emisor, quien le aprueba previa evaluación de su capacidad de endeudamiento y le otorga una línea de crédito. El consumidor, ahora es un

Cliente y al recibir la tarjeta es un Tarjeta Habiente, que le permite comprar bienes y servicios en todos aquellos comercios que aceptan esta tarjeta como forma de pago. Para hacer uso de su línea de crédito, el Cliente requiere una tarjeta de plástico, con el número identificador de la tarjeta y ciertos datos estampados, cinta magnética en el reverso y características de seguridad que pueden estar en el reverso y el anverso.

Los Emisores, son requeridos como resultado de su asociación con las operadoras internacionales de tarjetas de crédito como por ejemplo: Visa, Master Card o American Express, para que cumplan con ciertos requisitos específicos para cada marca, en la preparación de las tarjetas con el fin de que estas sean aceptadas en todas partes. Por otro lado, los comerciantes acuerdan con una institución financiera, que en adelante llamaremos Adquirente, la aceptación de las tarjetas de crédito como forma de pago. Con este fin, el comercio, abre una cuenta con el Banco Emisor. Este acuerdo permite que el comerciante venda sus productos y servicios a los clientes portadores de las tarjetas. La aceptación de las tarjetas implica en la mayoría de casos, el envío de transacciones electrónicas, a través de un punto de venta, al operador de la tarjeta el cual tiene comunicación con el emisor de la tarjeta y solicita la autorización de pago. El Banco Emisor revisa la cuenta del cliente para verificar su conformidad y responde aprobando o negando la operación. Esta respuesta la recibe el adquiriente a través del punto de venta. La aprobación implica que el banco emisor acuerda reembolsar el monto de la compra al adquirente, quien a su vez lo depositará en la cuenta del comerciante. El depósito en la cuenta es realizado al final del día, el punto de venta envía al banco emisor el resumen de las ventas efectuadas a través de un proceso por lotes. También es posible realizarlo mediante resúmenes de venta que el adquiriente llena manualmente y que deposita en las ventanillas del banco, a este resumen anexa los comprobantes firmados por cada uno de sus tarjetas habientes. En la siguiente figura se muestra el resumen del flujo expuesto.

2.2. Fraudes con Tarjeta de Crédito

El fraude con tarjeta de crédito implica el uso ilegal de la información o de la tarjeta de crédito física de una persona con el propósito de realizar compras o extraer fondos de su línea de crédito.

En el Perú existe un promedio de 450 usuarios de tarjetas de crédito que denuncian

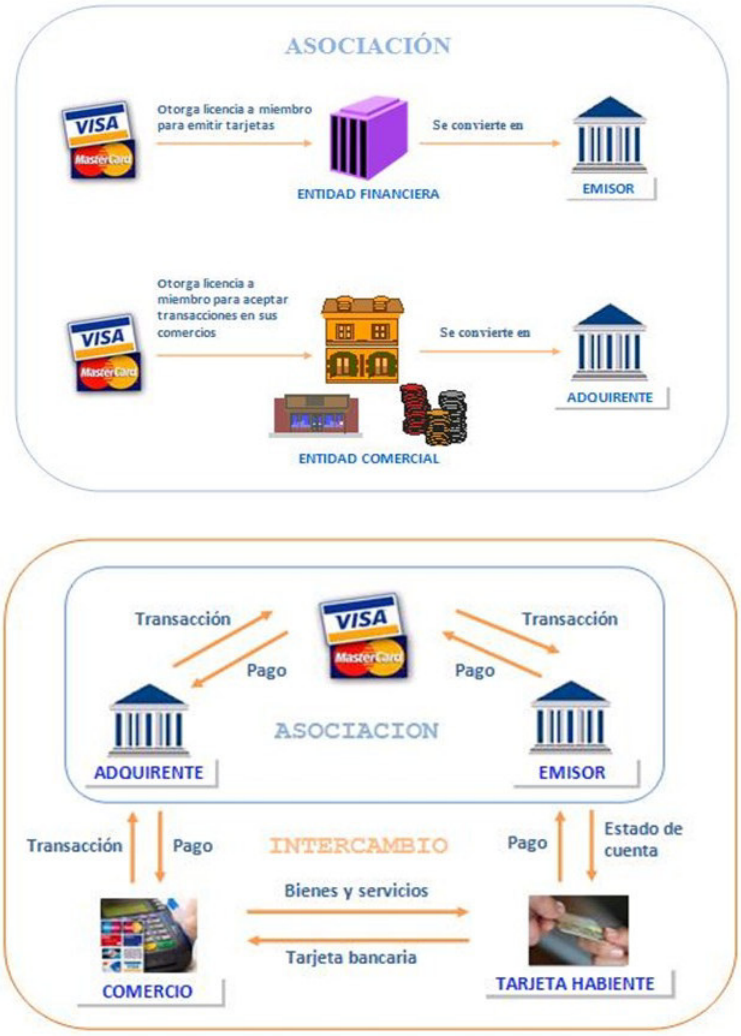

Figura 2.1: Flujo transaccional de una tarjeta de crédito

mensualmente fraude electrónico por consumos no realizados mediante transacciones financieras, cobros indebidos y reporte no justificado a centrales de riesgo [REF01]. Debido a la popularidad de las compras en línea, los delincuentes ya no necesitan una tarjeta de crédito física. Con el nombre del titular, el número de la tarjeta de crédito y la fecha de vencimiento es suficiente.

Los tipos de fraude más comunes con tarjetas de crédito son:

Virus troyano

Los virus troyanos son programas malintencionados capaces de alojarse en la pc de un cliente y permitir el acceso a usuarios externos, a través de una red local o de In-

ternet, con el fin de recabar información. Suele ser un programa alojado dentro de una aplicación, una imagen, un archivo de música u otro elemento de apariencia inocente, que se instala en el sistema al ejecutar el archivo que lo contiene. Una vez instalado parece realizar una función útil (aunque cierto tipo de troyanos permanecen ocultos y por tal motivo los antivirus o anti troyanos no los eliminan) pero internamente realiza otras tareas de las que el usuario no es consciente.

Cambio de tarjeta

Los estafadores al ver que un cliente va a retirar dinero en un cajero automático se le acercan, ofreciéndole ayuda o haciéndole alguna pregunta, confundiéndole y así logrando cambiar su tarjeta por otra del mismo banco pero que ya es inservible (Lo más probable de otra víctima) luego de haberle cambiado la tarjeta se encargan de observar la clave secreta y de esta manera obtienen los datos y realizan los retiros de dinero de su cuenta.

Duplicado de tarjeta o Skimming

La oficina del Defensor del Cliente Financiero (DCF) define al Skimming como una modalidad delictiva que consiste en la lectura no autorizada y en el almacenamiento de la información contenida en la banda magnética de las tarjetas bancarias, mediante la utilización de dispositivos electrónicos que brindan libre acceso a estos medios de pago[Vega 2009]. De esta manera el Skimming ha pasado a formar parte del universo de delitos de alta tecnología, en los que para perpetrarlos se hace uso de sofisticados equipos electrónicos y software especializado para perjudicar a las víctimas realizando consumos fraudulentos. Se puede emplear una micro cámara inalámbrica alimentada por una pequeña batería de 9v DC, o un simple teléfono celular con capacidad de tomar fotografías o filmar, con conexión a una computadora portátil (laptop) o hasta una lectora de bandas magnética (skimmer).

Phishing

Es aquel que se hace pasar por un correo electrónico legítimo de una organización, adjuntando enlaces o links a páginas falsas donde se solicita información confidencial

que puede ser utilizada para cometer algún tipo de fraude.

Pharming

Consiste en re direccionar al usuario hacia un URL fraudulento sin que éste se entere. Cuando un usuario trata de acceder a un URL, la dirección fraudulenta lo lleva hacia sitio fraudulento donde se le presenta una pantalla (similar a la original) en la cual ingresa su user y password. El sitio fraudulento responde que hay error en user y/o password y que se debe intentar de nuevo. Cuando el usuario reintenta, es direccionado al sitio legítimo.

Vshing

El cliente recibe una llamada telefónica, un e-mail o un mensaje de texto a su teléfono móvil, en el cual se le solicita llamar a un Sistema interactivo de voz (IVR) falso, a través del cual se capturan los datos de los clientes.

Ingeniería Social

En sí misma, la ingeniería social es la práctica de obtener información confidencial a través de la manipulación de usuarios legítimos. Con esta técnica, el ingeniero social se aprovecha de la tendencia natural del hombre a confiar en la gente, engañarles para romper los procedimientos normales de seguridad y manipularles para realizar acciones o divulgar información sensible. Otras técnicas usadas por el ingeniero social son las de apelar a la vanidad, la autoridad y la curiosidad de la gente. En general, se está de acuerdo en que los usuarios son el eslabón débil en seguridad y esto es aprovechado por la ingeniería social.

2.3. Descubrimiento de conocimiento en Bases de Datos

El descrubrimiento de conocimiento en Bases de Datos (KDD - por sus siglas en inglés: Knowledge Discovery in Databases), es el proceso no trivial de identificar pa-

trones válidos, novedosos, potencialmente útiles y, en última instancia comprensibles a partir grandes bases de datos [Fayyad et al., 1996].

La capacidad de generar y recolectar datos, debido al gran poder de procesamiento de los computadores así como a su bajo costo de almacenamiento, ha tenido un enorme crecimiento en los últimos años. Sin embargo, dentro de estos datos existe una gran cantidad de información oculta y de gran importancia para la toma de decisiones, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (DataMining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo el descubrimiento del conocimiento [Fayyad et al., 1996].

Las principales fases que integran el proceso de KDD se muestran en la figura 2.2 y son:

- Selección. Selección de los datos relevantes para el análisis (son obtenidos de la base de datos).
- Preproceso. Limpieza de datos, estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.
- Transformación. Conversión de datos en un modelo analítico, donde los datos se transforman o consolidan en formas apropiadas para la minería.
- Minería de Datos o Data Mining. Tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos para extraer conocimiento de los datos.
- Interpretación y evaluación. Identificación de patrones representativos del conocimiento.
- Conocimiento. Aplicación del conocimiento descubierto.

2.4. Minería de Datos

La Minería de Datos o Data Mining es una etapa del KDD y consiste en un conjunto de técnicas de múltiples disciplinas tales como: tecnología de bases de datos,

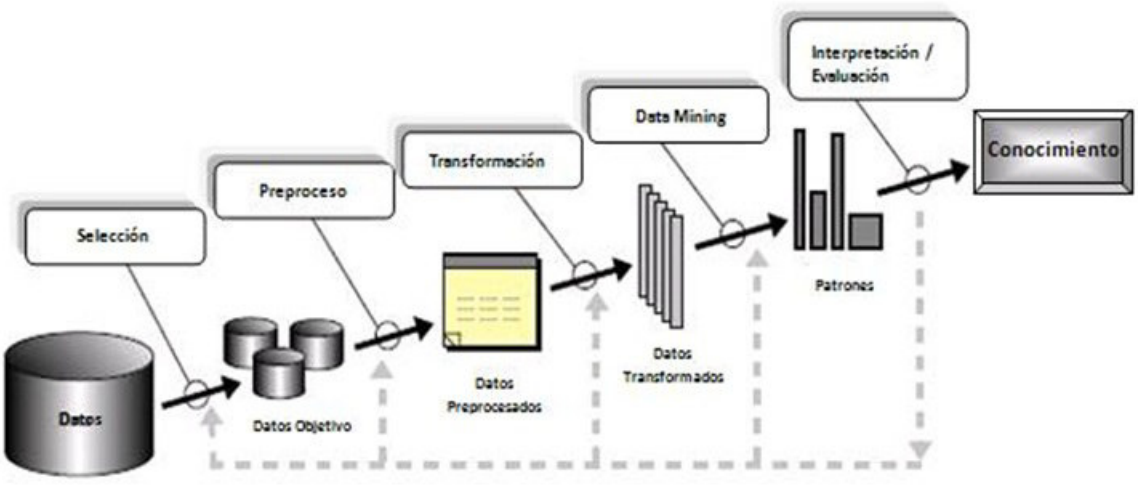

Figura 2.2: Proceso de KDD

estadística, aprendizaje, reconocimiento de patrones, redes neuronales, visualización de datos, obtención de información, procesamiento de imágenes y de señales, y análisis de datos [Fayyad et al 1996]. La idea de Minería de datos no es nueva, pues desde los años sesenta los estadísticos manejaban términos como data fishing, data mining o data archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A inicios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de data mining y KDD. [Vallejos06].

Dadas bases de datos de suficiente tamaño y calidad, la Minería de Datos puede generar nuevas oportunidades de negocios al proveer las siguientes capacidades:

- Predicción automatizada de tendencias y comportamientos. Automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, pueden ser contestadas directa y rápidamente desde los datos. Un ejemplo de ello son los segmentos de población que probablemente respondan similarmente a eventos dados.
- Descubrimiento automatizado de modelos previamente desconocidos. Las herramientas de Data Mining analizan las bases de datos e identifican modelos o estructuras ocultas en ellas. Otros problemas de descubrimiento de modelos incluye detectar patrones fuera de serie.

Para realizar estas tareas, la minería de datos se basa en diversos métodos. Según [Fayyad et al, 1996] la minería de datos utiliza los siguientes métodos:

- Clasificación: Se debe obtener un modelo que permita asignar datos sin un patrón conocido a unos patrones conocidos, para ello se utilizan datos conocidos y clasificados, de tal manera que el modelo pueda aprender las reglas de clasificación.

- Regresión: Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable.

- Clustering: Consiste en realizar una búsqueda de patrones en una base datos, con la particularidad de que no existe un conocimiento previo acerca de los patrones, siendo esta la principal diferencia con la clasificación. Los patrones se obtienen directamente de las bases de datos de entrada utilizando medidas de similaridad.

- Resumen: Consiste en métodos para encontrar una descripción compacta para un subconjunto de datos. Un ejemplo simple sería la tabulación de las desviaciones media y estándar para todos los campos. Métodos más sofisticados implican la derivación de las normas de resumen (Agrawal et al. 1996), técnicas de visualización múltiple, y el descubrimiento de las relaciones funcionales entre las variables (Zembowicz y Zytkow 1996). Técnicas de resumen a menudo se aplican al análisis exploratorio de datos interactiva y de generación automatizada de reportes.

- Modelados de dependencia existe en dos niveles: (1) el nivel estructural del modelo especifica (a menudo en forma gráfica), que son localmente variables dependientes el uno del otro. (2) El nivel cuantitativo del modelo especifica los puntos fuertes de las dependencias utilizando algunos números de escala. Por ejemplo, las redes de la dependencia probabilística uso independencia condicional para especificar el aspecto estructural del modelo y las probabilidades o correlaciones para especificar los puntos fuertes de las dependencias (Glymour et al. 1987; Heckerman 1996). Las redes de dependencia probabilística son cada vez más la búsqueda de aplicaciones en áreas tan diversas como el desarrollo de sistemas expertos médicos probabilística de bases de datos, recuperación de información, y el modelado del genoma humano.

2.5. Clustering

El Clustering consiste en la división de datos en grupos de objetos similares llamados Clústers [Mitra y Acharya 2006]. Los objetos son agrupados basándose en el principio de maximización de similitud dentro de los clusters y minimización de similitud entre clusters diferentes. No existe conocimiento previo a cerca de cómo deben conformarse los grupos, es por ese motivo que al clustering también se le considera como un técnica de aprendizaje no supervisado [Mitra y Acharya 2006].

El Clustering es una de las técnicas más útiles para descubrir conocimiento oculto en un conjunto de datos. En la actualidad el análisis de Clustering en Minería de Datos cumple un rol muy importante en una amplia variedad de áreas tales como: reconocimiento de patrones, análisis de datos espaciales, procesamiento de imágenes, cómputo y multimedia, análisis médico, economía, bioinformática y biometría principalmente [Hernandez, 2006]. Esto permite considerar el análisis de Clustering como una de las mejores técnicas para obtener conocimiento y realizar exploraciones en los datos.

Los datos a agrupar se encuentran representados por vectores reales, y la similitud se calcula en base a alguna medida que toma en cuenta los atributos de los datos. Por tanto el problema del clustering se puede formular de la siguiente manera: dado un conjunto de datos $X = (x_1, x_2, ..., x_n)$ en donde xi es un vector real, se deben encontrar K subconjuntos no vacios de X. Estos subconjuntos: $C_1, C_2, ..., C_k$ a los cuales se les denominará clusters, deben contener elementos similares y los grupos deben ser diferentes entre sí. Entonces, es importante para el proceso de clustering, la medida de la similitud a emplear, estas se revisarán a mayor detalle en las siguientes secciones.

2.5.1. Medidas de similitud

La medida de similitud seleccionada depende de las escalas de medida. Se pueden agrupar observaciones según la similitud expresada en términos de una distancia. Si se agrupan variables, es habitual utilizar como medida de similitud los coeficientes de correlación en valor absoluto, también llamados coeficientes de asociación. Para variables categóricas existen también criterios basados en la posesión o no de los

atributos (tablas de presencia-ausencia).

Basadas en la distancia

El concepto de distancia entre datos, representados por un vector real, permite interpretar geométricamente estos datos como puntos de un espacio métrico. La noción de similitud entre dos datos u objetos, representada por los vectores x_i y x_j de un conjunto $X \epsilon R^D$, se caracteriza por una función distancia entre estos vectores, donde cada vector de $X \epsilon R^D$ es D-dimensional, siendo $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$.

La función $d : X \times X \rightarrow R$ denota la medida de la distancia entre dos datos. Se dice que dos datos x_i y x_j son similares cuando la distancia entre los vectores es cercana a 0. La función de distancia tiene las siguientes propiedades [Chavez et al., 2001]:

- $\forall x_i, x_j \epsilon X, d(x_i, x_j) \geq 0$, positividad.
- $\forall x_i, x_j \epsilon X, d(x_i, x_j) = d(x_i, x_j)$, simétrica.
- $\forall x_i \epsilon X, d(x_i, x_j) = 0$, reflexiva.
- $\forall x_i, x_j \epsilon X, x_i \neq x_j \rightarrow d(x_i, x_j) > 0$, positividad estricta.
- $\forall x_i, x_j, x_k \epsilon X, d(x_i, x_j) \leq d(x_i, x_k) + d(x_j, x_k)$, desigualdad de triángulos.

Entre las medidas de distancia más conocidas se tienen [Marin, 2006]:

Distancia Euclidiana

Se utiliza con frecuencia para evaluar la similitud de objetos y obtiene buenos resultados cuando los objetos dentro de los clusters se encuentran de manera compacta y los clusters se encuentran separados unos de otros.

Dados dos objetos X_1 y X_2 medidos según dos variables x_1 y x_2, la distancia euclídea entre ambos es:

$$d_{I_1 I_2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2} \tag{2.1}$$

Con más dimensiones (o variables que se miden) es equivalente a:

$$d_{I_1 I_2} = \sqrt{\sum_{k=1}^{p} (x_{1k} - x_{2k})^2} \tag{2.2}$$

Uno de los inconvenientes que presenta la distancia euclidiana es cuando algunas características que tienen un amplio rango de valores predominan sobre las demás. Este problema puede mejorarse si se normalizan todos los valores a rangos de valores comunes o usando esquemas de ponderación. Otro de los inconvenientes de la distancia euclídea es que la correlación lineal entre las características de los objetos puede distorsionar la medida de la distancia [Vicente07].

Distancia de Minkowski

Se representa de la siguiente manera:

$$d_{I_i I_j} = \left[\sqrt{\sum_{k} |(x_{ik} - x_{jk})|^m} \right]^{\frac{1}{m}} \tag{2.3}$$

Donde $m \epsilon N$, siendo N todos los números naturales.

De esta distancia derivan 2 casos particulares:

Distancia de ciudad o de Manhattan ($m = 1$)

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{in} - x_{jn}| \tag{2.4}$$

Distancia Euclídea ($m = 2$)

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{in} - x_{jn}|^2} \tag{2.5}$$

Distancia de Mahalanobis

Es una medida de distancia introducida por Mahalanobis en 1936. Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales. Se diferencia de la distancia euclídea en que tiene en cuenta la correlación entre las variables aleatorias.

Se define como:

$$d^2_{I_i I_j} = (x_i - x_j)' W^{-1} (x_i - x_j)$$

Siendo W la matriz de covarianzas entre las variables. De este modo, las variables se ponderan según el grado de relación que exista entre ellas, es decir, si están más o menos correlacionadas. Si la correlación es nula y las variables están estandarizadas, se obtiene la distancia euclidiana.

Coeficientes de asociación

Los coeficientes de asociación son usados para calcular la similitud entre vectores binarios. Para dos datos o objetos representadas por vectores binarios x_i y x_j , el cálculo del coeficiente de asociación estará basado en el número de atributos coincidentes de una entidad en relación a otra. Lung et al. [Lung et al., 2004] clasifica a este tipo de coeficientes como cualitativos, debido a que calculan la similitud basado en la ausencia o presencia de atributos. Según Wiggerts [Wiggerts, 1997], son cuatro casos de asociación entre las entidades respecto al número de sus atributos:

- Presentes en ambas entidades,
- Presentes en x_i pero no en x_j,
- Presentes en x_j pero no en $x_i(c)$ y
- No presente en ambos

Si se denota por 1 binario como presencia de un atributo en una entidad, y por 0 la ausencia, es preferible relacionar la ocurrencia de esos atributos por una tabla definida como:

		x_j	
		1	0
x_i	1	a	b
	0	c	d

Por ejemplo, sean dos objetos x_i y x_j, descritos a través de dos vectores binarios $x_i = (0, 1, 0, 1, 1, 1)$ y $x_j = (0, 1, 1, 0, 1, 0)$, respectivamente. Entonces, $a = 2$ porque

los atributos presentes $(1-1)$ están en la segunda y quinta posición de ambos vectores. El valor de $b = 2$ porque los atributos cuarto y sexto están en x_i pero no en x_j, caso $(1 - 0)$. Así, se observan que $c = 1$, para $(0 - 1)$ y $d = 1$ para $(0 - 0)$.

Existen diversos métodos para calcular los coeficientes de asociación; ellos se diferencian en la relevancia que le dan a las coincidencias entre ambos vectores. Los principales métodos para el cálculo de coeficientes entre dos vectores x_i y x_j, usados en [Saeed et al., 2003; Wiggerts, 1997; Lung et al., 2004], son:

- Coeficiente de **Jaccard**: $S_j(x_i, x_j) = a/(a + b + c)$
- Coeficiente **Simple**: $S_s(x_i, x_j) = (a + d)/(a + b + c + d)$
- Coeficiente de **Sorensen**: $S_r(x_i, x_j) = 2a/(2a + b + c)$

Se observa que el coeficiente de Jaccard y Sorensen considera relevantes las relación $1 - 1$, pero no las relación $0 - 0$ ya que estas indican la ausencia de atributos. El coeficiente Simple, considera relevantes tanto las relaciones $1 - 1$, como las $0 - 0$.

En Patel et al. [1992], se propone una medida de similaridad en función de producto y norma de vectores binarios x_i y x_j, la cual también es usada para calcular la similaridad entre documentos por métodos de recuperación de información. La medida esta dada por la siguiente expresión:

$$S_1(x_i, x_j) = \frac{x_i \times x_j}{\|x_i\|\|x_j\|}.$$

En el mismo trabajo, se extiende esta función de medida a vectores no binarios, cuyos atributos expresan la frecuencia de ocurrencia de cierta característica. Por ejemplo, si $x_i = (x_{i1}, \ldots, x_{in})$ y $x_j = (x_{j1}, \ldots, x_{jn})$ representan a dos entidades de software (i.e. programas), entonces los valores de x_i y x_j pueden expresar la cantidad de veces que datos tipo T_i son declarados en dichos programas. La función de la medida extendida envuelve producto interno de vectores,

$$S_2(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|}.$$

2.5.2. Análisis de Clustering

El análisis de clustering, parte de un conjunto de datos u objetos cada uno de los cuales está caracterizado por varias variables. A partir de dicha información se

trata de obtener grupos de objetos, de tal manera que los objetos que pertenecen a un grupo sean muy homogéneos entre sí y, por otra parte, la heterogeneidad entre los distintos grupos sea muy elevada. Expresado en términos de variabilidad hablaríamos de minimizar la variabilidad dentro de los grupos para al mismo tiempo maximizar la variabilidad entre los distintos grupos.

$C_1 \dots C_i \dots C_k$
$O_1\ x_1{}^1 \dots x_i{}^1 \dots x_k{}^1$
.....
$O_j\ x_1{}^j \dots x_i{}^j \dots x_k{}^j$
.....
$O_n\ x_1{}^N \dots x_i{}^N \dots x_k{}^N$

Figura 2.3: Casos de partida para el análisis de clustering [Hernandez, 2006]

Estos métodos de análisis del clustering varían de acuerdo a los métodos aplicados [Vicente07]. Desde el punto de vista de la asignación de los objetos a los clusters, según Kearns [Kearns et al., 1997], los métodos que obtienen una solución al problema del clustering se dividen en dos tipos: el hard clustering y el soft clustering.

- El *hard clustering*, asume que los objetos deben ser asignados a uno y solo uno de los clusters, como consecuencia, los clusters encontrados son particiones de X, por lo tanto tendremos que: $C_i \cap C_j = \emptyset$, $\forall i, j = 1, \dots, K$ tal que $i \neq j$. El criterio de optimización en este caso es la minimización del error de la suma de los cuadrados de la distancia euclidiana entre los objetos y la media de clusters al que pertenecen. La heurística más conocida y usada es el algoritmo K-Means [Jain et al., 1999][Peña et al., 1999].

- El *soft clustering*, asume que cada objeto tiene un valor de membresía con respecto a cada cluster $C_i, i = 1, \dots, K$. Al contrario de las técnicas de hard clustering, los clusters encontrados no son particiones del conjunto de patrones observados; dentro de los algoritmos más conocidos se encuentran el Fuzzy C-Means [Bezdek, 1981] y Expectation Maximization [Dempster et al., 1977].

Referente a la organización de los objetos que forman los clusters, Jain et al [Jain et al., 1999] considera a los métodos particionales y los métodos jerárquicos. Estos a su vez pueden clasificarse a su vez en aglomerativos y divisivos. La diferencia entre estos

métodos radica en la forma de operar de los algoritmos. Los métodos aglomerativos comienzan con cada objeto como un cluster independiente, luego sucesivamente se unen entre ellos y los clusters más cercanos conforman un nuevo cluster, el proceso continúa hasta que es alcanzado algún criterio de parada. Por el contrario, los métodos divisivos parten de un conjunto de objetos como un único cluster y luego se divide sucesivamente en nuevos clusters hasta que es alcanzado algún criterio de parada. Esta clasificación será explicada con mayor detalle en el capítulo 3.

2.5.3. Características de los algoritmos de clustering

Las características deseables de la mayoría de los algoritmos de clustering son las siguientes:

Escalabilidad. La mayoría de los algoritmos de clustering trabajan de manera apropiada con un número pequeño de observaciones (hasta 200 aproximadamente), mientras que se necesita una gran escalabilidad para realizar agrupamiento de datos en bases con millones de observaciones.

Habilidad para trabajar con distintos tipos de atributos. Muchos algoritmos se han diseñado para trabajar sólo con datos numéricos, mientras que en una gran cantidad de ocasiones, es necesario trabajar con atributos asociados a tipos numéricos, binarios, discretos y alfanuméricos.

Descubrimiento de clusters con formas arbitrarias. La mayoría de los algoritmos de clustering se basan en la distancia euclidiana, lo que tiende a encontrar clusters todos con forma (circular) y densidad similares. Es importante diseñar algoritmos que puedan establecer clusters de formas arbitrarias.

Requerimientos mínimos en el conocimiento del dominio para determinar los parámetros de entrada. La herramienta no debería solicitarle al usuario que introduzca la cantidad de clases que quiere considerar, ya que dichos parámetros en muchas ocasiones no son fáciles de determinar, y esto haría que sea difícil controlar la calidad del algoritmo.

Habilidad para tratar con datos ruidosos. La mayoría de las BD contienen datos con comportamiento extraño, datos faltantes, desconocidos o erróneos. Algunos algoritmos de clustering son sensibles a tales datos y pueden derivarlos a clusters de baja

calidad.

Insensibilidad al orden de las observaciones de entrada. Algunos algoritmos son sensibles al orden en que se consideran las observaciones. Por ejemplo, para un mismo conjunto de datos, dependiendo del orden en que se analicen, los clusters devueltos pueden ser diferentes. Es importante entonces que el algoritmo sea insensible al orden de los datos, y que el conjunto de clusters devuelto sea siempre el mismo.

Alta dimensionalidad. Una BD o DW (DataWarehouse) puede contener varias dimensiones o atributos, por lo que es bueno que un algoritmo de clustering pueda trabajar de manera eficiente y correcta no solo en repositorios con pocos 30 atributos, sino también en repositorios con un alto espacio dimensional, o gran cantidad de atributos.

Clustering basado en restricciones. Es un gran desafío el agrupar los datos teniendo en cuenta no sólo el comportamiento, sino también que satisfagan ciertas restricciones.

Interpretación y uso. Los usuarios esperan que los resultados del clustering sean comprensibles, fáciles de interpretar y de utilizar.

Con estas características, se busca diseñar algoritmos más flexibles que sean capaces de manipular una gran variedad de requerimientos de acuerdo a las necesidades de los usuarios.

2.6. Detección de Anomalías

La detección de anomalías o outliers se refiere al problema de encontrar datos o grupos de datos que no se ajustan al comportamiento esperado. La detección de outliers tiene uso en una amplia variedad de aplicaciones como:

- Detección de fraudes en tarjeta de crédito, seguros o servicios de salud.
- Detección de intrusos para la seguridad informática.
- Detección de fallos en la seguridad de sistemas críticos y la vigilancia militar para las actividades del enemigo.

La detección de Outliers fue estudiada por la comunidad de estadística en el siglo XIX. Con el tiempo, una variedad de técnicas se han desarrollado para detectar valores anómalos en las comunidades de investigación.

2.6.1. Definición de Outlier

Son los patrones en los datos que no se ajustan a un concepto bien definido de comportamiento normal [Chandola, 2007].

En la siguiente Figura se muestra los Outliers en un conjunto de datos bi-dimensional. En esta figura los datos tienen dos regiones normales: N_1 y N_2, así como puntos que están lo suficientemente lejos de estas regiones: O_1, O_2 y O_3, estos puntos son los Outliers.

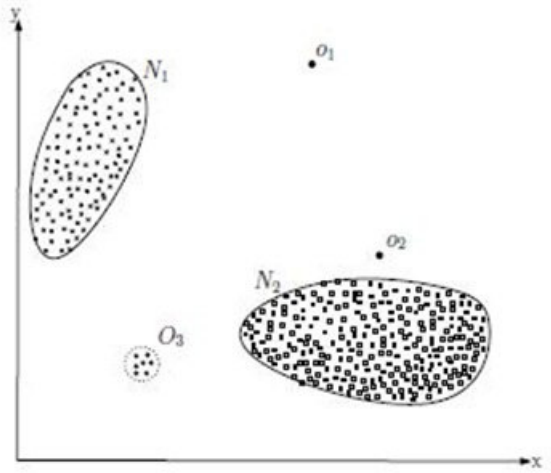

Figura 2.4: Conjunto de datos bi-dimensional [Chandola, 2007]

2.6.2. Desafíos de los Outliers

Un desafío clave en la detección de Outliers es que se trata de explorar el espacio invisible. Un acercamiento directo será definir una región que representa un comportamiento normal y declarar cualquier observación en los datos que no pertenece a esta región normal como Outlier. Pero varios factores hacen este acercamiento, al parecer simple, muy desafiante:

- La definición de una región normal que abarque cada comportamiento normal

posible es muy difícil.

- El límite entre el comportamiento normal y anormal no es a menudo exacto. Así una observación considerada como comportamiento anormal, cerca del límite, puede ser realmente normal y viceversa.
- La noción exacta de un Outlier es diferente para diversos dominios del uso. Cada dominio del uso impone un sistema de requisitos y de apremios que dan lugar a una formulación específica del problema para la detección de Outlier.
- La disponibilidad de los datos etiquetados para la validación es a menudo un tema importante mientras que desarrolla una técnica de detección del Outlier.
- En varios casos, los Outliers son el resultado de acciones malintencionadas, los adversarios malévolos se adaptan para hacer que las observaciones anormales aparezcan como normales, de tal modo que la tarea de definir un comportamiento normal se hace más difícil.
- Los datos contienen a menudo ruido que es similar a los outliers reales y por lo tanto es difícil de distinguir y quitar.
- Muchas veces el comportamiento normal sigue evolucionando y la noción actual de la conducta normal puede no ser suficientemente representativa en el futuro.

Debido a los desafíos mencionados, el problema de detección Outliers en su forma más general, no es fácil de resolver. De hecho, la mayor parte de las técnicas existentes para la detección del mismo simplifican el problema centrándose en una formulación específica.

La formulación es inducida por varios factores tales como la naturaleza de los datos, disponibilidad de etiquetado datos, tipo de Outliers que se detectarán, etc. A menudo, estos factores son determinados por el dominio del uso en el cual la técnica debe ser aplicada.

2.6.3. Tipos de Outlier

Se pueden clasificar en tres categorías basadas en su composición y su relación al resto de los datos:

Outlier Puntual

En un determinado conjunto de instancias de datos, una instancia externa, llamada también periférica, se denomina como un punto de valor anormal (Outlier). Este es el tipo más sencillo y es el foco de la mayoría de los actuales sistemas de detección.

En la Figura 2.4 las áreas N_1 y N_2 representan la región normal de datos. Los puntos O_1 y O_2, así como los puntos en la región de O_3 se encuentran fuera de los límites de la región normal y por lo tanto son Outliers.

Como ejemplo de la vida real vamos a considerar detección de fraude en el uso de la tarjeta de crédito, el conjunto de datos correspondería a las transacciones con tarjetas de crédito que realiza una persona. Supongamos para este ejemplo que los datos se definen mediante una sola característica: cantidad de dinero gastado. Una transacción con una cantidad gastada muy alta en comparación con el rango normal de los gastos de esa persona será un Outlier.

Outlier Contextual

También denominado Condicional [Song et al. 2007] surge de la ocurrencia de un caso individual de los datos en un determinado contexto. Los Outlier Contextual se definen con respecto a un contexto, el cual se induce por la estructura del conjunto de datos y es especificado como parte de la formulación del problema. Al igual que los Outlier puntuales, estos Outliers también son casos individuales de los datos. La diferencia es que no puede denominarse como Outlier en otro contexto diferente al especificado inicialmente.

Satisface 2 propiedades:

1. Cada instancia de datos se define mediante dos conjuntos de atributos, a saber:

- Atributos contextuales y
- Atributos de comportamiento.

Por ejemplo, en conjuntos de datos espaciales, la longitud y la latitud de una localización son las cualidades del contexto, en datos de la serie cronológica el tiempo

es una cualidad del contexto que determina la posición de un caso respecto a la secuencia entera. Las cualidades del comportamiento definen las características no contextuales de un caso. Por ejemplo, en un conjunto de datos espaciales que describen la precipitación media del mundo entero, la cantidad de precipitación en cualquier localización es una cualidad del comportamiento.

2. El comportamiento de un Outlier es resuelto usando los valores de los atributos de comportamiento dentro de un contexto específico. Una instancia de datos puede ser un outlier contextual en un contexto dado, pero una instancia de datos idénticos (en términos de atributos de comportamiento) puede considerarse normales en un contexto diferente.

Esta propiedad es clave en la identificación de los atributos contextuales y de comportamiento para una técnica de detección de Outliers.

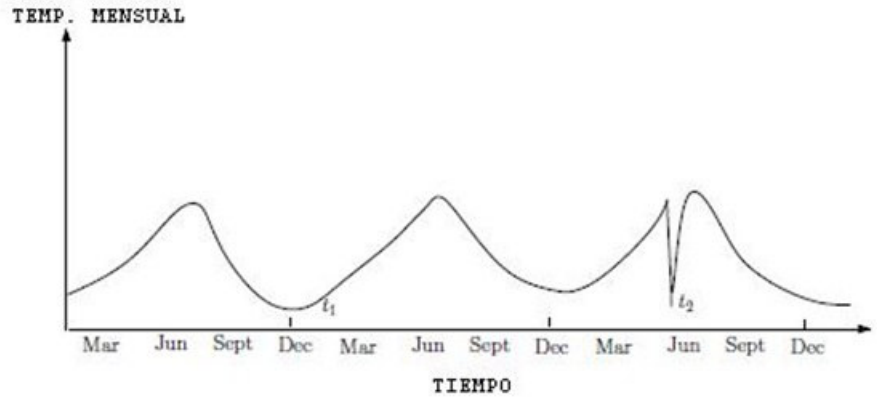

Figura 2.5: Outlier contextual t2 en una serie Tiempo-Temperatura [Chandola, 2007]

La Figura anterior muestra un ejemplo de ello para una serie Tiempo-Temperatura que muestra la temperatura mensual de un área durante los últimos años. Una temperatura de 35F puede ser normal durante el invierno (en el tiempo t1) en ese lugar, pero el mismo valor durante el verano (en el tiempo t2) sería un outlier.

Un ejemplo similar puede encontrarse en el dominio de la detección de fraudes de tarjeta de crédito. Un atributo de contexto en la tarjeta de crédito puede ser el lugar de compra. Un hombre puede gastar alrededor de $ 10 en una estación de gas, mientras que su pareja podría gastar alrededor de $ 200 en una tienda de joyas. Una nueva operación de $ 200 en la gasolinera sería considerada como un outlier contextual, ya que no se ajusta a la conducta normal de la persona en el contexto de la estación de gas (a pesar de que la misma cantidad que gastan en la tienda de

joyería se considerarán normal).

La aplicación de técnicas de detección de outliers contextuales, está determinada por la significación de los outliers en el dominio de aplicación de destino. Otro factor clave es la disponibilidad de los atributos contextuales. En varios casos la definición de un contexto es sencillo, y por lo tanto la aplicación de la técnica de detección tiene sentido. En otros casos, la definición de un contexto no es fácil, por lo que se dificulta a aplicación de estas técnicas.

Outlier Colectivo

Estos Outliers ocurren en datos donde casos de datos individuales están relacionados y una colección de casos de datos relacionados es periférica en lo que concierne al juego de datos entero. Los casos de datos individuales en un colectivo outlier no pueden ser Outliers en forma separada, pero su presencia conjunta como una colección es anómala.

La siguiente figura ilustra un ejemplo que muestra una salida de electrocardiograma humano. La línea ampliada plana denota un Outlier porque el mismo valor bajo existe para un modo anormal mucho tiempo. Note que aquel valor bajo, por sí mismo, no es un Outlier.

Figura 2.6: Outlier contextual t2 en una serie Tiempo-Temperatura [Chandola, 2007]

Cabe señalar que, si bien un Outlier Puntual puede ocurrir en cualquier conjunto de datos, el Outlier Colectivo sólo puede ocurrir en conjuntos de datos relacionados. En contraste, la aparición de Outliers contextuales depende de la disponibilidad de los atributos de contexto en los datos. Un Outlier Colectivo también puede ser Outlier contextual si se analiza con respecto a un contexto. Así pues, un problema de detección de Outlier puntual o Colectivo puede ser transformado en un problema de detección de Outlier Contextual mediante la incorporación de la información del contexto.

En el desarrollo de la tesina, nos basaremos en el tipo de outlier puntual, cuya

aplicación será de gran utilidad en la detección de fraude para las tarjetas de crédito.

Capítulo 3

Estado del Arte

3.1. Técnicas de Clustering

Los algoritmos de clustering pueden clasificarse en función de las siguientes características:

- El tipo de dato que manejan (numérico, categórico y/o mixto).
- El criterio utilizado para medir la similitud entre los puntos.
- Los conceptos y técnicas de clustering empleadas (ejemplo: lógica difusa, estadísticas).

Una clasificación general divide los algoritmos en:

- Clustering jerárquico.
- Clustering basado en particiones.
- Clustering basado en grid.

A continuación se describirán las técnicas de clustering más representativas en la minería de Datos.

3.1.1. Clustering Jerárquico

Un método jerárquico crea una descomposición jerárquica de un conjunto de datos, formando un dendograma, el cual divide recursivamente el conjunto de datos en conjuntos cada vez más pequeños. Un dendrograma es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de clusters. Los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud/disimilitud entre los objetos. Este método se divide en 2 subclases:

El método aglomerativo o bottom-up

Empieza con un grupo por cada objeto y une los grupos más parecidos hasta llegar a un solo grupo u otro criterio de paro.

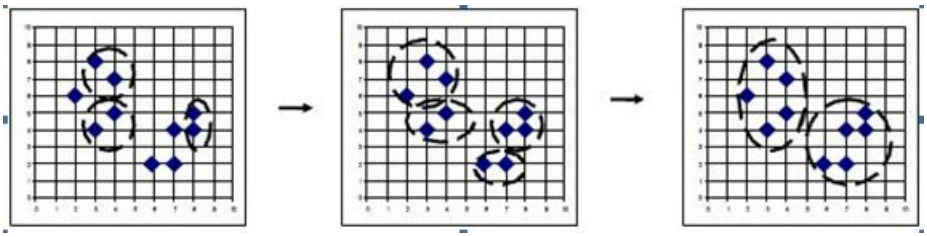

Figura 3.1: Método Jerárquico Aglomerativo
[Alonso, 2008]

El método divisorio o top-down

Empieza con un solo grupo y lo divide en grupos más pequeños hasta llegar a grupos de un solo elemento u otro criterio de paro.

Entre los algoritmos más conocidos para esta clasificación se encuentran [Hernandez, 2006]:

- CURE (Clustering Using Representatives)
- CHAMALEON
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchical)

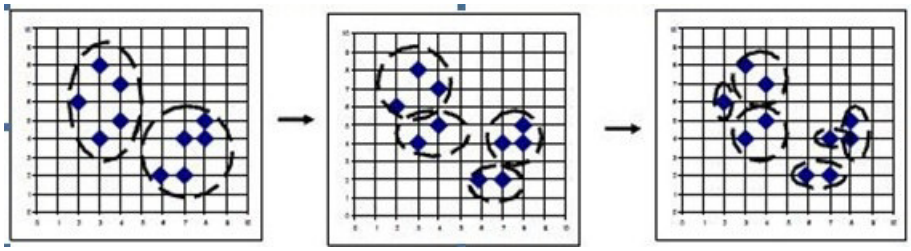

Figura 3.2: Método Jerárquico Divisorio
[Alonso, 2008]

- ROCK (Robust Clustering algorithm using links)
- AGNES (AGlomerative NESting)
- DIANA (Divisive ANAlysis)

3.1.2. Clustering basado en particiones

Este método construye k particiones de los datos, donde cada partición representa un grupo o cluster. Cada grupo tiene al menos un elemento y cada elemento pertenece a un solo grupo. Estos métodos, crean una partición inicial e iteran hasta un criterio deparo. Los más populares son k-medias y k-medianas.

3.2. Métodos de Clustering

3.2.1. Hard Clustering

Los métodos que comprende el hard clustering son exhaustivos y exclusivos. Es decir, cada uno de los objetos contenidos en el conjunto $X \in R^D$ debe ser asignado a solo uno de los clusters, en consecuencia los clusters generados son particiones del total de objetos observados. Entonces, para los clusters C_i y C_j,

$$C_i \bigcap C_j = \emptyset, \text{ para todo } i, j = 1, \ldots, K, \text{ e } i \neq j.$$

Formulado de esta manera, el clustering es un problema NP-Difícil [Brucker, 1978][Garey y Johnson, 1979]. Encontrar una solución exacta al problema requiere la evaluación de una cantidad extremadamente grande de configuraciones de clusters para determinar

cual es la mejor.Por tanto, el uso de algoritmos enumerativos para obtener la solución exacta al problema del clustering resulta impráctico, y por este motivo es bien justificado el uso de métodos heurísticos y metaheurísticos para encontrar soluciones factibles del problema.

Para la evaluación de los clusters se debe usar una función objetivo que mida la similaridad de los objetos dentro de cada cluster. Como los objetos de un cluster son similares cuando las distancias entre ellos es mínima; esto permite formular la función objetivo f, como:

$$f = \sum_{j=1}^{K} \sum_{x_i \in C_j} d(x_i, \bar{x}_j)^2; \tag{3.1}$$

donde $\bar{x}_j$, conocido como elemento representativo del cluster, es la media de los elementos del cluster C_j:

$$\bar{x}_j = \frac{1}{|C_j|} \sum x_i \in C_j, \tag{3.2}$$

y corresponde al centro del cluster.

Algoritmo K-Means

El algoritmo K-Means [Forgy, 1965][McQueen, 1967] es una de las heurísticas mas sencillas y comúnmente utilizadas para resolver el problema de clustering. El algoritmo clasifica de una manera sencilla los objetos de X en K clusters conocidos a priori. La idea general del algoritmo es obtener los centros iniciales y formar clusters asociando los objetos de X a los centros más cercanos. Luego de que se han asociado todos los elementos de X, se recalculan los centros. Si no hay variación con respecto a los centros anteriores entonces el algoritmo termina en caso contrario se repite el proceso de asociación hasta que no haya variación en los centros, o se cumpla algún otro criterio de parada.

Específicamente, el algoritmo inicia seleccionando aleatoriamente K diferentes objetos del conjunto X; tales objetos serán los centros iniciales denotados por: $\{\bar{x}_j\}_{j=1,\ldots,K}$. La asociación del objeto $x_i \in X$ con el centro más cercano $\bar{x}_j$ del cluster C_j es dada si $d(x_i, \bar{x}_j) < d(x_i, \bar{x}_p)$ para todo $j, p = 1, \ldots, K$ y $j \neq p$. Luego de asignar cada uno de los objetos $x_i \in X$ con el centro mas cercano, los centros son recalculados usando la expresión (3.2). El proceso se repite hasta que los centros no varíen, o se llegue a alguna otra condición de parada, como alcanzar un número determinado de

iteraciones o cuando las reasignaciones de los objetos sea muy poca [Jain et al., 1999]. El pseudocódigo del algoritmo K-Means se presenta a continuación:

Algoritmo 3.1: K-Means

entrada: $X = \{x_1, ..., x_n\}, K$

1 **inicio**
2 Seleccionar aleatoriamente e X centros iniciales $\{\bar{x}_i\}_{i=1,...,K}$;
3 **para cada** $x_i \in X$ **hacer**
4 Asociar x_i con el centro mas cercano: $C_j = C_j \bigcup \{x_i\}$, si $d(x_i, \bar{x}_j) < d(x_i, \bar{x}_p) \forall j, p = 1, ..., K$ y $j \neq p$;
5 **fin**
6 Calcular los centros $\bar{x}_i^* = \frac{1}{|C_i|} \sum_{j=1}^{|C_i|} x_j$, para $x_j \in C_i$;
7 **si** *no hay mas reasignaciones:* $\bar{x}_i^* = \bar{x}_i$, $\forall i$ **entonces**
8 Parar;
9 **sino**
10 considerar $\bar{x}_i^*$ como nuevo centro $\bar{x}_i$, e ir al paso 3;
11 **fin**
12 **fin**

El algoritmo K-Means no es exento de inconvenientes, los principales, según Peña et al. [Peña et al., 1999], son:

- Es sensible a la inicialización. Es decir, la solución final es dependendiente de los centros iniciales.
- Como otros algoritmos heurísticos, el K-Means es un algoritmo determinístico que converge rápidamente a óptimos locales. El algoritmo minimiza el criterio de optimización dado en (3.1), aunque no obtiene una configuración óptima de los clusters.
- Se debe tener conocimiento previo del valor de K. Este inconveniente se maneja teniendo a K como un parámetro de entrada del algoritmo.

En la literatura, se han propuesto diversas adaptaciones del algoritmo K-Means para mejorar los inconvenientes descritos. En [Bradley y Fayyad, 1998] se propone una mejora para aliviar el inconveniente de la sensibilidad a la inicialización. Primero se

obtienen soluciones del K-Means sobre pequeñas muestras inicializando las aleatoriamente. Cada una de las soluciones obtenidas es utilizada como posible inicialización al aplicar el K-Means sobre la unión de todas las muestras. Los centros de la mejor solución obtenida son usados para inicializar el algoritmo K-Means sobre todo el conjunto de patrones.

En [Peña et al., 1999] se discuten cuatro métodos para la inicialización del algoritmo K-Means: de manera completamente aleatoria, la propuesta de Forgy [Forgy, 1965], la propuesta de McQueen [McQueen, 1967] y la propuesta de Kauffman y Rousseeuw [Kaufman y Rousseeuw, 1990]. Las tres primeras propuestas son de alguna manera aleatorias, sólo el algoritmo propuesto por Kauffman y Rouseeuw es un algoritmo heurístico que identifica los K objetos más representativos que prometen tener a su alrededor una gran cantidades de patrones. La inicialización completamente aleatoria y la propuesta de Kauffman y Rouseeuw proporcionan una mejor inicialización para el algoritmo K-Means que el resto de métodos, haciéndolo más robusto; aunque, el segundo método converge más rápidamente que la inicialización aleatoria.

En [Likas et al., 2003] se propone una adaptación del algoritmo K-Means para hacer una búsqueda determinística de los clusters. El método esta basado en la idea que la solución óptima para encontrar los K clusters puede ser obtenida haciendo una serie de búsquedas locales basadas en el algoritmo K-Means. En cada búsqueda local de K clusters, $K-1$ centros se encuentran inicializados en sus posiciones óptimas correspondiente a la solución al problema con $K-1$ clusters. El K-ésimo cluster es encontrado inicializando el K-ésimo centro en múltiples objetos de X y eligiendo la solución óptima. Es decir, el proceso debe encontrar las j-ésimas soluciones al problema de clustering para $j = 1, \ldots, K$. Como la solución óptima cuando $j = 1$ es conocida, entonces es posible encontrar incrementalmente la solución para el problema de los K clusters.

En [Elkan, 2003] se propone una aceleración para el algoritmo K-Means, la cual consiste en evitar el cálculo de la distancia de los objetos a algunos centros, incluso al centro del cluster al cual se encuentra asignado. Elkan demuestra dos lemas basados en el teorema de desigualdad de triángulos, el cual afirma que dados tres puntos x, y y z la distancia de uno de los lados, es menor o igual que la suma de las distancias de los otros dos, es decir, la propiedad 5 de la función de distancia.

Lema 1: Dados un objeto x y dos centros $\bar{x}_a$ y $\bar{x}_b$:

$$\text{si } d(\bar{x}_a, \bar{x}_b) \geq 2d(x, \bar{x}_a) \text{ entonces } d(x, \bar{x}_b) \geq d(x, x_a).$$

Este primer lema permite evitar el cálculo de la distancia del objeto x, asignado a un cluster con centro $\bar{x}_a$, a otro centro $\bar{x}_b$, cuando la distancia entre los centros $\bar{x}_a$ y $\bar{x}_b$ es mayor a dos veces la distancia entre x y $\bar{x}_a$. La distancia $d(x, \bar{x}_a)$ tampoco sera necesaria calcularla si se tiene un límite superior u de tal forma que $u \geq d(x, \bar{x}_a)$ y se cumple que $d(\bar{x}_a, \bar{x}_b) \geq u$.

Lema 2: Dados un objeto x y dos centros $\bar{x}_b$ y $\bar{x}_b^{'}$, entonces:

$$d(x, \bar{x}_b) \geq Max\{0, d(x, \bar{x}_b^{'}) - d(\bar{x}_b^{'}, \bar{x}_b)\}.$$

Este lema permite tener un límite inferior de la distancia de un punto x a un centro x_b, de tal manera que, si $\bar{x}_b^{'}$ es el centro del cluster en una iteración previa y $l_b^{'}$ un limite inferior tal que $d(x, \bar{x}_b^{'}) \geq l_b^{'}$, podemos decir que el limite inferior en la iteración actual es: $d(x, x_b) \geq Max\{0, d(x, \bar{x}_b{'}) - d(\bar{x}_b, \bar{x}_b^{'})\}$, por lo tanto $d(x, \bar{x}_b) \geq Max\{0, l_b^{'} - d(\bar{x}_b, \bar{x}_b^{'})\} = l_b$. El algoritmo propuesto por Elkan evita el cálculo de las distancias de un punto a los centros haciendo uso de límites superiores e inferiores de la siguiente manera: si $u \geq d(x, \bar{x}_a)$ es un limite superior de $d(x, \bar{x}_a)$, siendo $\bar{x}_a$ el centro del cluster al que se encuentra asignado x, y $l_b \leq d(x, \bar{x}_b)$ el limite inferior de la distancia de x y el centro $\bar{x}_b$, si se cumple que $u \leq l_b$ entonces $d(x, \bar{x}_a) \leq u \leq l_b \leq d(x, \bar{x}_b)$ y no es necesario calcular las distancias $d(x, \bar{x}_a)$ y $d(x, \bar{x}_b)$, caso contrario se deben calcular las distancias. Esta reducción de cantidad de cálculos de la distancia entre objetos y centros, tiene como consecuencia un mejor tiempo de respuesta del algoritmo K-Means.

Métodos Metaheurísticos

Los algoritmos Genéticos [Goldberg, 1989] para el problema del clustering han sido propuestos por Murthy y Chowdhurry [Murthy y Chowdhury, 1996], Maulik y Bandyopadhyay [Maulik y Bandyopadhyay, 2000][Bandyopadhyay y Maulik, 2002], Pacheco y Valencia [Pacheco y Valencia, 2003], entre otros. Estas metaheurísticas mejoran el inconveniente de la convergencia a óptimos locales del algoritmo K-Means. En la propuesta de Murthy y Chowdhurry se codifican las soluciones en cromosomas de longitud igual al número de elementos de X, por lo que el método está limitado por

el número de elementos de X. Las otras propuestas codifican los cromosomas con los valores de los centros de los clusters. En este caso, un cromosoma está compuesto por un vector $a = (\bar{x}_1, \ldots, \bar{x}_K)$, donde un $\bar{x}_i$ es un centro de un cluster, su implementación requiere de menos recursos y el método es factible para cualquier número de elementos de X. Es decir, la propuesta de Maulik y Bandyopadhyay es parecida con la propuesta de Pacheco y Valencia en el aspecto de codificación de cromosomas, pero varían en cuanto a los operadores de *cruzamiento* y *mutación*, y que el primero utiliza el algoritmo K-Means para refinar la solución en cada generación de la población.

Los Algoritmos Meméticos [Moscato, 1989] han sido propuestos para el hard clustering por Pacheco y Valencia [Pacheco y Valencia, 2003] y Merz [Merz, 2003]. De manera similar a los Algoritmos Genéticos, los Algoritmos Meméticos utilizan poblaciones de soluciones denominadas memes, que se van recombinando generación tras generación en búsqueda de un óptimo. La diferencia radica en que cada meme es obtenido por un algoritmo de búsqueda local, por tanto la búsqueda se realiza en un espacio de soluciones correspondientes a óptimos locales. Merz propone el uso del algoritmo K-Means para la generación de los óptimo locales, mientras que Pacheco y Valencia realizan experimentos con diversos algoritmos de búsqueda local, tales como HK-Means y J-Means [Hansen y Mladenovic, 2001]. Tanto Merz como Pacheco y Valencia codifican los memes con los centros de los clusters; es decir, un meme esta representado por un vector $a = (\bar{x}_1, \ldots, \bar{x}_K)$ de centros obtenidos con un algoritmo de búsqueda local. Merz utiliza un operador de cruzamiento que promete eliminar los centros del padre a que no tienen un centro cercano en el padre b, mientras que Pacheco y Valencia hacen un ordenamiento previo de los padres a y b de acuerdo a la cercanía de los centros y luego hacen el cruzamiento.

La metaheurística GRASP (Greedy Randomized Adaptative Search Procedure) [Feo y Resende, 1995] ha sido propuesta para el problema del clustering por Cano et al. [Cano et al., 2002]. La metaheurística GRASP es un proceso de multiarranque compuesta por dos fases: una fase de construcción en la que se generan buenas soluciones de manera aleatoria en base a un algoritmo goloso; y una fase de búsqueda local, en que se busca una mejoría de las soluciones obtenidas en la fase de construcción. Cano et al. utilizan en la fase de construcción una adaptación del algoritmo de Kauffman y Rouseeuw [Kaufman y Rousseeuw, 1990] para encontrar los centros iniciales, y en la fase de búsqueda local usa el algoritmo K-Means para mejorar la solución encontrada en la fase previa. El método demuestra ser superior al algoritmo

K-Means en las colecciones de datos usadas.

3.2.2. Soft Clustering

Los métodos de hard clustering obtienen clusters que corresponden a particiones del conjunto de objetos analizados X, debido a que dichos métodos asocian cada objeto a solo uno de los clusters. Los métodos de soft clustering extienden esta noción asociando cada objeto con todos los clusters, regulando esta asociación por una función de membresía dada por:

$$m(C_j \mid x_i),$$

que define la porción del objeto x_i pertenece al cluster C_j. Dicha función de membresía debe cumplir con las siguiente condiciones:

$$m(C_j \mid x_i) \geq 0 \text{ y } \sum_{j=1}^{K} m(C_j \mid x_i) = 1;$$

cuando es $m(C_j \mid x_i) = 0$ quiere decir que x_i no pertenece al cluster C_j, y cuando es $m(C_j \mid x_i) = 1$ quiere decir que x_i pertenece completamente al cluster C_j. La forma de calcular el valor de $m(C_j \mid x_i)$ depende del método a usar.

A continuación se presentan los dos métodos más usados para el soft clusterig: Fuzzy K-Means y Expectation Maximization.

Fuzzy K-Means

El algoritmo conocido como Fuzzy K-Means [Dunn, 1974][Bezdek, 1981] soluciona el problema del clustering asignando a cada uno de los objetos $x_i \in X$ a uno o más clusters de acuerdo a una función de membresía. Esta heurística tiene un mejor comportamiento que el algoritmo K-Means en el inconveniente de la convergencia a óptimos locales, aunque sin dejar de tener el mismo problema.

El algoritmo Fuzzy K-Means optimiza el error cuadrático de las medias de los clusters, pero a diferencia del K-Means, asigna a cada elemento x_i un valor de membresía denotado por u_{ij} que indica el grado de pertenencia del objeto x_i con respecto al cluster C_j. Dunn [Dunn, 1974] propone la minimización del error cuadrático ponderado

con el valor de membresía de cada elemento, de la siguiente forma:

$$\sum_{i=1}^{N}\sum_{j=1}^{K}(u_{ij})^2 d(x_i, \bar{x}_j)^2, \tag{3.3}$$

donde los valores para u_{ij} se encuentran en el intervalo $[0, 1]$ y $\sum_{j=1}^{K} u_{ij} = 1$ para todo $x_i \in X$. Los valores para cada u_{ij} se encuentran en una matriz $U_{N\times K}$.

Bezdek [Bezdek, 1981] propone una generalización del criterio dado en (3.3), asignando un valor variable al exponente de u_{ij}, el cual le llama coeficiente fuzzy. El criterio de optimización propuesto por Bezdek esta dado por:

$$\sum_{i=1}^{N}\sum_{j=1}^{K}(u_{ij})^m d(x_i, \bar{x}_j)^2, \tag{3.4}$$

donde el valor de m está en el intervalo $[1, \infty)$.

El clustering es llevado acabo a través de un proceso iterativo que optimiza la función objetivo dado en (3.4). El proceso actualiza el valor de membresía u_{ij} y los centros de los clusters $\bar{x}_j$ como:

$$\bar{x}_j = \frac{\sum_{i=1}^{N}(u_{ij})^m x_i}{\sum_{i=1}^{N}(u_{ij})^m}, \tag{3.5}$$

donde

$$u_{ij} = \frac{1}{\sum_{l=1}^{K}\left(\frac{d(x_i,\bar{x}_j)^2}{d(x_i,x_l)^2}\right)^{\frac{2}{m-1}}} \tag{3.6}$$

El proceso iterativo termina cuando $Max_{ij}\{|u_{ij}^{l+1} - u_{ij}^{l}|\} < \epsilon$, donde ϵ se le conoce como criterio de parada cuyo valor se encuentra en el intervalo $[0, 1]$, y l es el número de iteraciones. El algoritmo formalizado en pseudo-código se presenta a continuación:

Algoritmo 3.2: Fuzzy K-Means

entrada: $X = \{x_1, ..., x_n\}, K$

inicio
  Inicializar los clusters aleatoriamente y la matriz de membresía $U^{(0)} = [u_{ij}]$;
  para $l = 1, 2, 3, \ldots$ **hacer**
    Calcular los centros $\bar{x}_j, j = 1, \ldots, K$ usando (3.5) y $U^{(l-1)}$;
    Calcular $U = U^{(l-1)}$ usando (3.6) y $\bar{x}_j$;
    si $\{|u_{ij}^{l+1} - u_{ij}^{l}|\} < \epsilon$ **entonces**
      Parar;
    fin
  fin
fin

Al igual que el K-Means, el algoritmo empieza seleccionando K clusters aleatorios de X, e inicializa los valores de la matriz U de forma aleatoria. En cada iteración se van actualizando los valores de los centros y los valores de la matriz hasta que el algoritmo converge a un mínimo local de la expresión (3.4). Uno de los principales problemas al implementar el algoritmo es la elección del exponente fuzzy, en [Bezdek, 1981] se propone varias técnicas para su elección.

Expectation Maximization

Una manera de solucionar el problema del clustering es aproximando los clusters a modelos estadísticos. Esto consiste en seleccionar un modelo estadístico y ajustar iterativamente clusters a dicho modelo. En la práctica, cada cluster puede ser representado matemáticamente por una distribución de probabilidad, tal como la Gaussiana (continua) o Poisson (discreta). De esta manera, los clusters son representados por un modelo llamado finite mixtures (mixtura de distribuciones), donde cada distribución de probabilidad corresponde con un cluster y se refiere a ella como un componente del modelo. Por ejemplo, si tenemos dos funciones de distribución P_0 y P_1 correspondientes a dos clusters, un objeto $x_i \in X$ y $P_0(x_i) \geq P_1(x_i)$, entonces, el algoritmo asigna parcialmente el objeto x_i al cluster con distribución P_0, con un valor de $P_0(x_i)/(P_0(x_i) + P_1(x_i))$ y el resto del valor de x_i al cluster con distribución P_1.

El algoritmo Expectation Maximization(EM)[Dempster et al., 1977], es una heurística que asume que la generación de los cluster de objetos está dada por una mixtura de distribuciones de probabilidad Normal o Gaussiana, el objetivo es identificar los parámetros de la función de distribución de probabilidad para cada cluster. El proceso consiste en dos pasos iterativos de optimización, el paso "E" estima las probabilidades de que un patrón pertenezca a un determinado cluster $P(x_i \mid C_j)$ y el paso "M" encuentra una aproximación a la mixtura de parámetros de función de distribución de probabilidad que definen los clusters. Estos dos pasos se repiten iterativamente hasta que se encuentran los parámetros que maximiza el siguiente criterio de optimización:

$$-\sum_{i=1}^{N} log\left(\sum_{j=1}^{K} P(x_i/C_j)P(C_j)\right) \tag{3.7}$$

Debido a los fundamentos probabilísticos del algoritmo Expectation-Maximization, este método alivia el inconveniente que presenta el algoritmo K-Means de identificar clusters de forma esférica. También es robusto respecto del ruido que puedan contener los datos observados. El algoritmo EM también presenta algunas desventajas, al igual que el K-Means, una inicialización pobre puede llevar a que el algoritmo converga lentamente y se deben conocer de a priori el número de clusters.

3.3. Detección de Anomalías (Outlier)

La meta principal en la detección de Anomalías, es encontrar objetos que sean diferentes de los demás (Outlier), estos objetos anómalos tienen valores de atributos con una desviación significativa respecto a los valores típicos esperados.

Aunque los Outlier son frecuentemente tratados como ruido en muchas operaciones, para propósitos de detección de fraudes son una herramienta valiosa para encontrar comportamientos atípicos en las operaciones que un cliente realiza.

El análisis de conglomerados [Jain y Dubes 1988], es una máquina popular de técnicas de aprendizaje agrupados por instancias de datos similares. El Clustering es principalmente una técnica de supervisión si la agrupación semi-supervisada [Basu et al. 2004] también se ha explorado últimamente.

A pesar de que la agrupación y la detección de valores atípicos parecen ser fundamentalmente diferentes unos de otros, la agrupación de varias técnicas basadas en

la detección de valores atípicos han sido desarrolladas. Estas técnicas se basan en el supuesto fundamental de que los puntos de datos normales pertenecen a grupos grandes y densos, mientras que los valores extremos, o bien no pertenecen a ningún grupo o forman parte de grupos muy pequeños. El Clustering basado en técnicas de detección de valores atípicos puede ser ampliamente clasificado a lo largo de dos dimensiones:

¿Qué suponen las etiquetas?

En esta categoría las técnicas se pueden agrupar en técnicas de supervisión y no supervisión.

- **Las técnicas bajo supervisión** comprenden el uso normal de datos para generar agrupaciones que representan los modos normales de comportamiento de los datos [Marchette 1999; Wu y Zhang 2003; Vinueza y Grudic 2004].

  Cualquier instancia nueva es asignada a uno de los grupos, si ésta no pertenece a ninguno de los grupos comprendidos, se le denomina valor atípico, separa los datos normales de los valores extremos mediante la minería conjunto de elementos frecuentes. Los datos se dividen en segmentos basados en el tiempo. Para cada segmento, se generan conjuntos de elementos frecuentes. Todos los conjuntos de elementos que existen en más de un segmento se consideran normales. Todos los datos de los puntos correspondientes a la normalidad conjuntos de elementos frecuentes se utilizan para obtener los grupos de limpieza y utilizando la técnica de agrupación CoolCat [Barbara et al. 2002].

- **Técnicas de uso no supervisado** de un algoritmo de agrupamiento es conocido por agrupar datos y luego analizarlos independientmente respecto a los grupos.

¿Cómo son los valores anómalos detectados?

La mayor parte de la anterior agrupación basada en técnicas de detección de valores atípicos encontrando valores anómalos como el subproducto de un algoritmo de clustering [Impuestos y Duin 2001; Ester et al. 1996; Jain y Dubes 1988; Ng y Han 1994]. Así pues, cualquier dato que no es considerado en ningun grupo que se llama un valor atípico.

Varias agrupaciones se centran en las técnicas basadas en la detección de valores atípicos, en lugar de generar agrupaciones. El algoritmo CLAD [Mahoney et al. De 2003] se deriva de la anchura de los datos, tomando una muestra al azar y el cálculo de la distancia media entre los puntos más cercanos. Todos aquellos grupos cuya densidad es inferior a un umbral se declaran como totales "valores extremos", mientras que todos los grupos que están lejos de otros grupos se declaran como globales "valores atípicos". Una variante del algoritmo K-means clustering se utiliza para la detección de valores atípicos por Jiang et al. [2001] usando un enfoque similar.

La ventaja de las técnicas basadas en agrupaciones es que no tienen que ser supervisados. Además, las técnicas basadas en la agrupación son susceptibles de ser utilizados en un modo incremental, es decir, después de conocer los grupos, los nuevos puntos pueden ser alimentados en el sistema.

Una desventaja de las técnicas basadas en la agrupación es que son computacionalmente costosas, ya que implican el cálculo de las distancias en pares. Agrupación de ancho fijo es un algoritmo de aproximación [Epiel et al. 2002; Portnoy et al. 2001; Mahoney et al. 2003; He et al. 2003]. Un punto es asignado a un grupo cercano, si no existe tal grupo entonces se crea un nuevo grupo tomando como centro dicho punto. Luego, determinar qué grupos contienen los valores extremos en función de su densidad y la distancia de las otras categorías. Chaudhary et al. [2002] propone una técnica de detección de valores atípicos mediante árboles kd que proporcionan una compartimentación de los datos en tiempo lineal. Se aplica su técnica para detectar valores atípicos en conjuntos de datos astronómicos.

3.4. Trabajos relacionados

3.4.1. Definición de outliers basados en la distancia

Se define como un objeto que está en la distancia mínima d_{min} de la distancia porcentual de k respecto de los objetos en el conjunto de datos.

El problema es entonces encontrar la dmin adecuada y k.

Definición: Un punto x en un conjunto de datos es un caso aparte respecto a los parámetros k y d.

Para explicar la definición por ejemplo tomamos parámetro $k = 3$ y la distancia d como se muestra en la siguiente figura, éstos son los puntos X_i y X_j los cuales se definen como outliers, en el interior del círculo para cada punto, no se encuentran más de 3 otros puntos. Y x' es un inlier, porque se ha excedido el número de puntos dentro del círculo dado los parámetros k y d.

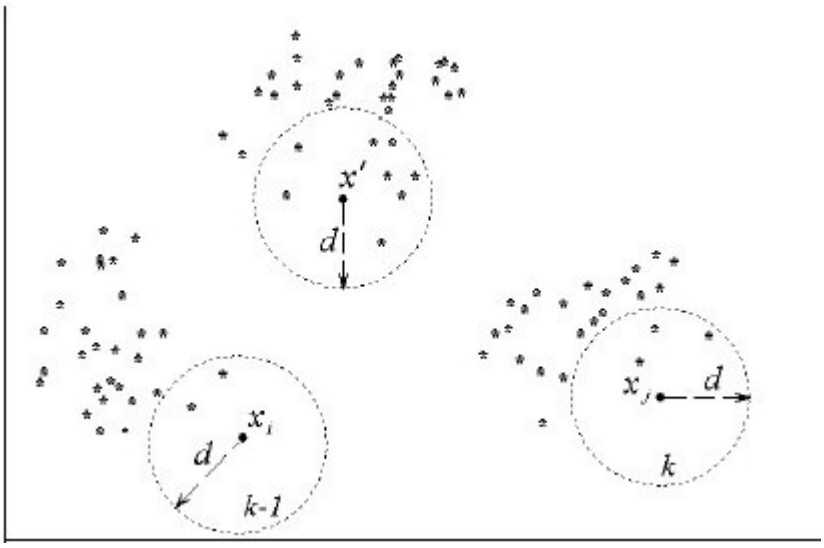

Figura 3.3: Definición de outliers por Knorr y Ng

Este enfoque no requiere ningún conocimiento a priori de las distribuciones de datos. Sin embargo, este enfoque basado en la distancia tiene ciertas deficiencias:

1. Se requiere que el usuario especifique una distancia d, lo que podría ser difícil determinar a priori.

2. No proporciona una clasificación para los outliers: por ejemplo, un punto con muy pocos puntos próximos a una distancia d puede considerarse en cierto sentido como un outlier.

Se hace cada vez más difícil de estimar el parámetro d con aumento de la dimensionalidad.

Así, si uno toma un poco pequeño radio d, entonces todos los puntos son los valores extremos. Si uno toma un radio d grande, entonces no tiene sentido es un caso aparte. Así, el usuario debe elegir D a un grado muy alto de precisión a fin de encontrar un modesto número de puntos que pueda definirse como Outliers.

En el siguiente gráfico, X_i y X_j son considerados como outliers, pero d_1 es demasiado grande, por lo tanto, dentro de los círculos hay demasiados puntos. En este caso definimos X_i y X_j como inliers incorrectos.

Por el contrario si el d_2 es demasiado pequeño y dentro de cada círculo se encuentran pocos puntos, luego x_l y x_k pueden considerados incorrectamente como outliers [Knorr y Ng].

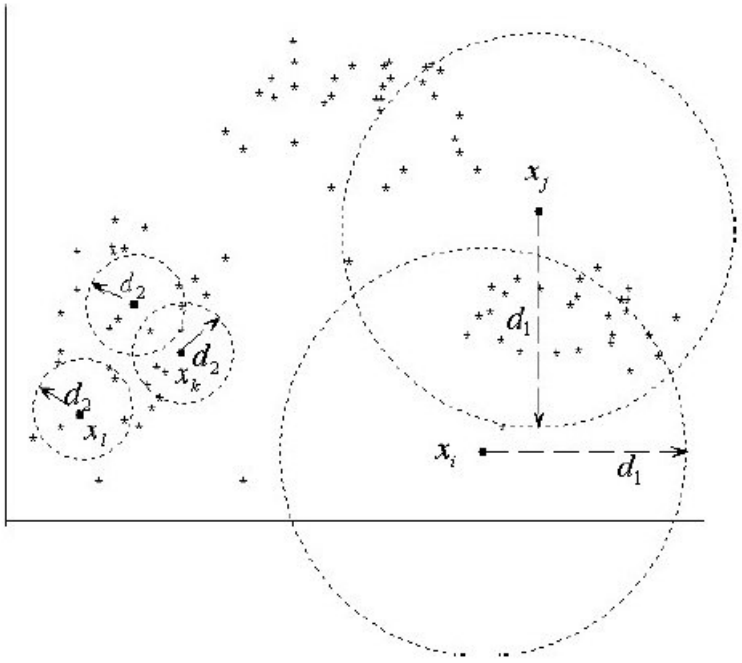

Figura 3.4: Definición de las deficiencias

Capítulo 4

Método Aplicado

En base a la revisión de las técnicas de Clustering, se aplicará un método basado en las técnicas predictivas, el cual permitirá encontrar los clusters y outliers de una población de datos, con la finalidad de determinar para el caso de estudio, las posibles ocurrencias de fraude.

Este método considera los siguientes pasos:

1. Ubicación de los mejores centros o clusters de datos, empleando para ello un algoritmo basado en particiones K-Means.
2. Ubicación del punto más lejano, con la finalidad de reafirmar los elementos de cada cluster y detectar los outliers, que para el caso expuesto, serían los posibles fraudes a considerar.
3. En el caso se obtuviesen clusters con pocos elementos, también se consideran como posibles outliers.

4.1. Ubicación de los mejores centros

En este paso se clasifican los patrones, de un población N de datos, en K clusters o aglomerados siendo K un dato previamente conocido. Esta clasificación se realiza con el algoritmo K-Means [Forgy, 1965][McQueen, 1967], el cual determina los puntos centrales de cada cluster al cual está asociado cada dato de la población.

Para ubicar los mejores centros se realizan los siguientes pasos:

1. Se indica el número de clusters (K).
2. Se determina el número de iteraciones (P) a realizar, para ubicar los mejores centros.
3. Se considera aleatoriamente K datos de la población N, como centros, para dar inicio al algoritmo K-Means, cuyo flujo se muestra en la figura 2.1 (algortimo).
4. Luego de cada iteración del algoritmo, se calcula el error cuadrático determinado como:

$$\Delta\iota_w = \sum_{1}^{N} d(X_i, C_j); \forall_i = 1, ..., N; j = 1, ..., K; w = 1, ..., P \tag{4.1}$$

   En la fórmula previa se calcula, por cada iteración la sumatoria de las distancias euclidianas de cada patrón a su respectivo centro del cluster.
5. Se comparan los errores cuadráticos entre la iteración previa y la actual, resultando la más optima aquella que presente el menor valor. Por cada iteración, calificada como óptima, se guarda la referencia de los centros hallados.
6. Al finalizar las P iteraciones, se obtiene el error cuadrático óptimo y los mejores centros, que serán el punto de partida para encontrar el punto más distante a los centros de cada cluster.

4.2. Ubicación del punto mas cercano

Tomando como base los centros iniciales, se procede con los siguientes pasos:

1. Por cada cluster, se determina la distancia promedio de cada patrón de datos a su cluster.

$$DIST_{PROM} = \frac{\sum_1^n (X_i, C_j)}{n}; \forall_i = 1, ..., n; j = 1, ..., K \tag{4.2}$$

Siendo n, el número de elementos de cada cluster j. Posteriormente se toma en cuenta la siguiente relación, cuyo valor al ser mayor a 1, se considera como punto lejano.

$$REL = \frac{d(X_i, C_j)}{DIST_{PROM_j}} \tag{4.3}$$

2. De cada cluster se obtiene el punto más lejano y se compara con los puntos lejanos de los K clusters, considerando el de mayor valor como Outlier.

3. Este punto lejano, Outlier, es eliminado de la población X_i procediendo a ejecutar nuevamente el algoritmo K-Means.

4. Se repite desde el paso 1, tantas veces como Outliers se determinen ubicar. El número de iteraciones es determinado por el experto del negocio.

Al ubicar los mejores centros, se reafirmar los elementos de cada cluster y se detectan los posibles Outliers de una población de datos(X).

4.3. Cluster pequeños

En el caso se determine clusters pequeños, con pocos elementos, estos deben ser revisados debido a ser posibles Outliers. Se toma como referencia lo determinado por el Método de las 2 fases: "Los cluster pequeños y los árboles con menor número de nodos, son seleccionados y considerados como Outliers"[Jiang01].

Capítulo 5

Implementación

Como primera instancia se valida la implementación del método aplicado, para lo cual se prueba con un Dataset bidimensional de 3000 registros. Se procederá a probar los resultados, comparando los cluster obtenidos, considerando como variables de entrada:

Número de iteraciones. Representa el número de veces que se ejecutará el algoritmo K-Means para obtener el mínimo error cuadrático, requerido en el primer paso del método aplicado.

K inicial. Indica el número de clusters, cuyos centros serán los datos de entrada para el 2do paso del método a implementar.

Outliers. Representa el número de puntos lejanos a eliminar del data set o población de datos.

5.1. Verificación del método con un data set predeterminado

5.1.1. Descripción del dataset

Se considera un dataset de 2 dimensiones, cuyo total de registros es 3000. El objetivo es validar el método propuesto con un dataset predeterminado, considerado en aplicaciones de otros métodos de Clustering, existentes a la fecha.

5.1.2. Validación del método propuesto

En este primer caso de verificación el K es conocido, por lo cual se validará el método aplicado haciendo variaciones del número de Outliers, obteniendo los elementos o clusters considerados como Outliers.

Valor de K	20
Outliers	
Caso	Nro. Outliers
1	2
2	3
3	5

Ubicación de los mejores centros

El primer paso consiste en obtener los mejores centros, para lo cual se obtienen los errores cuadráticos de cada iteración, como se muestra en la figura 5.1.

Al iterar 400 y 500 veces, se comprueba que los centros se estabilizan. Se considerará para este caso el valor de 400 como punto de partida para la siguiente fase del método propuesto.

Ubicación del punto más lejano

Para los 3 casos referentes al número de Outliers a detectar, se determinará los elementos a excluir del cluster correspondiente aplicando para ello los pasos del punto 4.2.

A continuación se mostrará los resultados obtenidos al ubicar los puntos más lejanos.

Caso 1: Detección de 2 outliers

Se itera el algoritmo tantas veces como número de Outliers se deseen obtener. En este caso se itera 2 veces, debido a que el número de Outliers considerado es 2, ver figura 5.2.

Los elementos eliminados de los cluster 11 y 13 se muestran en la figura 5.3, se debe tomar en cuenta que las posiciones se contabilizan desde la posición 0, por lo cual el rango de posiciones va desde 0 hasta 2999.

Caso 2: Detección de 3 outliers

Se itera el algoritmo 3 veces para encontrar los 3 Outliers determinados. Ver figura 5.4.

Los elementos eliminados de los cluster 4, 15 y 17 se muestran en la figura 5.5, se debe tomar en cuenta que las posiciones se contabilizan desde la posición 0, por lo cual el rango de posiciones va desde 0 hasta 2999.

Caso 3: Detección de 5 outliers

Se itera el algoritmo 5 veces para encontrar los 5 Outliers determinados. Ver figura 5.6.

Los elementos eliminados de los cluster 3,4,8,11 y 20 se muestran en la figura 5.7, se debe tomar en cuenta que las posiciones se contabilizan desde la posición 0, por lo cual el rango de posiciones va desde 0 hasta 2999.

Al ser eliminados los 5 puntos más lejanos, los centros de cada cluster se reajustan, consiguiendo que los elementos de cada cluster sean más homogéneos entre ellos. En el Gráfico 5.8 se muestra la distribución final de los 3000 elementos en los 20 clusters, así como los 5 Outliers detectados, siendo estos los marcados mediante un círculo.

Clusters pequeños

En este caso el número de elementos por cluster es similar, siendo la diferencia ente 1 y 5 elementos. Por lo tanto este paso del método no aplicaría.

Al validar el método aplicado, se puede concluir que los Outliers detectados se mantienen constantes desde el primer caso, lo cual confirma la consistencia del método propuesto.

5.2. Descripción del dataset de operaciones con tarjetas de crédito

El dataset considerado para el caso de estudio, es una muestra de 3000 operaciones reales, realizadas por los usuarios de tarjetas de crédito, en Lima-Perú. Las operaciones se han realizado con Nuevos Soles, restringiendo así el caso a validar con el método propuesto.

Asimismo, se consideran 2 tipos de operaciones:

- **Avance de efectivo**, disposición de efectivo en cajeros autorizados, con cargo a la línea de crédito. El valor identificador en el data set es 7 para este tipo de operación.
- **Compras**, utilizando como medio de pago la tarjeta de crédito, en diversos centros de comercio. El valor identificador en el data set es 5 para este tipo de operación.

Las dimensiones consideradas son:

- Tipo de operación
- Importe de la operación (Nuevos Soles)

5.2.1. Validación del método propuesto

Se consideran los siguientes valores iniciales, antes de la ejecución del método:

K inicial: 2

Outliers : 2

Ubicación de los mejores centros

El primer paso consiste en obtener los mejores centros, para lo cual se obtienen los errores cuadráticos de cada iteración, como se muestra en la figura 5.9.

Al iterar 2000 y 2500 veces, se comprueba que los centros se estabilizan. Se considerará para este caso el valor de 2000 como punto de partida para la siguiente fase del método propuesto.

Ubicación del punto más lejano

Siendo el número de Outliers a detectar igual a 2, se ejecuta 2 veces el algoritmo respectivo para obtener los 2 puntos o patrones a excluir del dataset. En la figura 5.10 se muestra el cluster que pierde 2 elementos al ser considerados como Outliers.

Los elementos eliminados del cluster 1 se muestran en la figura 5.11, se debe tomar en cuenta que las posiciones se contabilizan desde la posición 0, por lo cual el rango de posiciones va desde 0 hasta 2999.

Al ser eliminados los 2 puntos más lejanos, los centros de cada cluster se reajustan, consiguiendo que los elementos de cada cluster sean más semejantes. En la figura 5.12 se muestran la evolución de los centros.

Clusters pequeños

Para verificar si existen clusters pequeños se debe conocer el número de elementos de los clusters existentes. En la figura 5.13 se muestra el número de elementos por cada cluster, así como el rango de valores que contiene cada cluster, referente a la dimensión importe.

Se observa que el 2do cluster contiene sólo 1 elemento y es el importe más elevado del dataset, lo cual se confirma con el rango de valores del cluster 1. Por lo tanto se

concluye que la operación realizada por el monto de $S/.28900.79$, es un posible caso de fraude por uso de tarjeta de crédito.

En la figura 5.14 se muestran los 3 Outliers identificados con el método propuesto.

Nro. Iteraciones	Error Cuadrático	Mejores Centros		
		Centro	Dimensión	
			X	Y
100	1.408653	1	36177.364963	50878.328467
		2	54461.229729	42821.513513
		3	30826.253424	45588.582191
		4	51573.673202	49327.294117
		5	19925.423841	61146.205298
		6	35624.989304	59208.946524
		7	4770.297297	54722.594594
		8	36598.063492	46427.539682
		9	44782.503311	46320.695364
		10	56813.570469	35697.355704
		11	38537.613207	43952.660377
		12	60959.315436	46441.026845
		13	10325.342105	50846.026315
		14	28134.263358	58578.377862
		15	23495.155405	45438.837837
		16	58641.16	59852.84
		17	60331.403973	52416.397350
		18	35903.597315	37050.805369
		19	10595.533333	60619.72
		20	17043.573333	54549.933333
300	1.214633	1	58641.16	59852.84
		2	60959.315436	46441.026845
		3	36419.550335	58760.543624
		4	51573.673202	49327.294117
		5	44743.973856	46293.712418
		6	30868.574324	45602.141891
		7	60331.403973	52416.397350
		8	54461.229729	42821.513513
		9	19903.18	61170.266666
		10	23495.155405	45438.837837
		11	30783.091549	61002.507042
		12	4770.297297	54722.594594
		13	36164.862745	50580.098039
		14	38001.033557	44522.120805
		15	26783.647798	56977.786163
		16	10595.533333	60619.72
		17	17043.573333	54549.933333
		18	56813.570469	35697.355704
		19	10325.342105	50846.026315
		20	35903.597315	37050.805369
400	1.214625	1	26778.107594	56958.025316
		2	17043.573333	54549.933333
		3	4770.297297	54722.594594
		4	54461.229729	42821.513513
		5	51573.673202	49327.294117
		6	10325.342105	50846.026315
		7	36419.550335	58760.543624
		8	60331.403973	52416.397350
		9	35903.597315	37050.805369
		10	19903.18	61170.266666
		11	44743.973856	46293.712418
		12	58641.16	59852.84
		13	30761.244755	60996.195804
		14	38001.033557	44522.120805
		15	10595.533333	60619.72
		16	60959.315436	46441.026845
		17	30868.574324	45602.141891
		18	23495.155405	45438.837837
		19	36164.862745	50580.098039
		20	56813.570469	35697.355704
500	1.214625	1	60959.315436	46441.026845
		2	30868.574324	45602.141891
		3	10325.342105	50846.026315
		4	58641.16	59852.84
		5	17043.573333	54549.933333
		6	30761.244755	60996.195804
		7	51573.673202	49327.294117
		8	38001.033557	44522.120805
		9	23495.155405	45438.837837
		10	56813.570469	35697.355704
		11	26778.107594	56958.025316
		12	54461.229729	42821.513513
		13	19903.18	61170.266666
		14	10595.533333	60619.72
		15	35903.597315	37050.805369
		16	4770.297297	54722.594594
		17	60331.403973	52416.397350
		18	44743.973856	46293.712418
		19	36419.550335	58760.543624
		20	36164.862745	50580.098039

Figura 5.1: Ubicación de los mejores centros con el data set predeterminado

	Nro. de elementos por Cluster – Detección de Outliers		
Cluster	Inicial	Iteración 1	Iteración 2
1	153	153	153
2	149	149	149
3	149	149	149
4	158	158	158
5	149	149	149
6	150	150	150
7	153	153	153
8	150	150	150
9	148	148	148
10	150	150	150
11	148	148	147
12	153	153	153
13	148	147	147
14	149	149	149
15	151	151	151
16	150	150	150
17	148	148	148
18	152	152	152
19	143	143	143
20	149	149	149
TOTAL	3000	2999	2998

Figura 5.2: Elementos por cluster con 2 Outliers detectados

Outlier		Dimensión	
Cluster	Posición	X	Y
11	2808	0.0	54499.0
13	1127	28736.0	40792.0

Figura 5.3: Elementos eliminados de los clusters 11 y 13

	Nro. de elementos por Cluster – Detección de Outliers			
Cluster	Inicial	Iteración 1	Iteración 2	Iteración 3
1	150	150	150	150
2	149	149	149	149
3	149	149	149	149
4	151	151	151	150
5	150	150	150	150
6	149	149	149	149
7	149	149	149	149
8	152	152	152	152
9	149	149	149	149
10	143	143	143	143
11	158	158	158	158
12	153	153	153	153
13	153	153	153	153
14	150	150	150	150
15	148	147	147	147
16	150	150	150	150
17	148	148	147	147
18	148	148	148	148
19	153	153	153	153
20	148	148	148	148
TOTAL	3000	2999	2998	2997

Figura 5.4: Elementos por cluster con 3 Outliers detectados

Outlier		Dimensión	
Cluster	Posición	X	Y
4	524	65535.0	51609.0
15	1127	28736.0	40792.0
17	2808	0.0	54499.0

Figura 5.5: Elementos eliminados de los clusters 4, 15 y 17

	Nro. de elementos por Cluster – Detección de Outliers					
Cluster	Inicial	Iteración 1	Iteración 2	Iteración 3	Iteración 4	Iteración 5
1	153	153	153	153	153	153
2	158	158	158	158	158	158
3	149	149	149	149	148	148
4	151	151	151	150	150	150
5	150	150	150	150	150	150
6	149	149	149	149	149	149
7	150	150	150	150	150	150
8	148	148	147	147	147	147
9	153	153	153	153	153	153
10	153	153	153	153	153	153
11	148	147	147	147	147	147
12	149	149	149	149	149	149
13	143	143	143	143	143	143
14	149	149	149	149	149	149
15	152	152	152	152	152	152
16	150	150	150	150	150	150
17	148	148	148	148	148	148
18	150	150	150	150	150	150
19	148	148	148	148	148	148
20	149	149	149	149	149	148
TOTAL	3000	2999	2998	2997	2996	2995

Figura 5.6: Elementos por cluster con 5 Outliers detectados

Outlier		Dimensión	
Cluster	Posición	X	Y
3	1663	39884.0	40043.0
4	524	65535.0	51609.0
8	2808	0.0	54499.0
11	1127	28736.0	40792.0
20	1356	51782.0	34639.0

Figura 5.7: Elementos eliminados de los clusters 3, 4, 8, 11 y 20

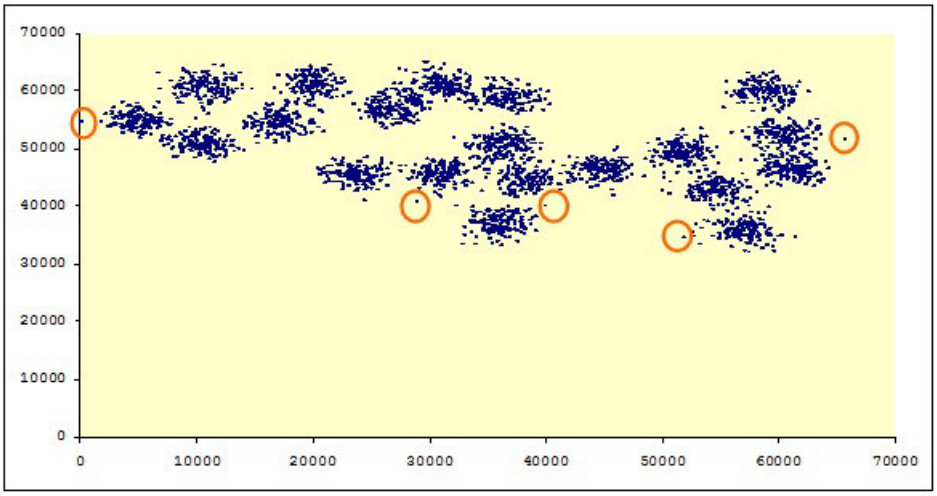

Figura 5.8: Distribución de los 3000 datos con los 5 Outliers detectados

Nro. Iteraciones	Error Cuadrático	Mejores Centros		
		Centro	Dimensión	
			X	Y
500	4.271786	1	5.552552	126.411161
		2	5.0	16229.543333
1000	4.271786	1	5.0	16229.543333
		2	5.552552	126.411161
2000	3.770191	1	5.552184	132.925005
		2	5.0	28900.79
2500	3.770191	1	5.0	28900.79
		2	5.552184	132.925005

Figura 5.9: Ubicación de los mejores centros

	Nro. de Elementos		
Cluster	Inicial	Sin Outlier 1	Sin Outlier 2
1	2999	2998	2997
2	1	1	1
Total	3000	2999	2998

Figura 5.10: Cluster con variación del número de elementos

Outlier		Dimensión	
Cluster	Posición	X	Y
1	2430	5.0	10787.84
1	17	5.0	9000.00

Figura 5.11: Elementos eliminados del cluster 1

Centro	Dimensión	
	X	Y
Inicial		
1	5.552184	132.925005
2	5.0	28900.79
Sin Outlier 1		
1	5.552368	129.370997
2	5.0	28900.79
Sin Outlier 2		
1	5.552552	126.411161
2	5.0	7711.781764

Figura 5.12: Evolución de los centros de cada cluster

	Rango de Importes		Nro. de Elementos
CLUSTER	Min	Max	
1	0.55	6000.00	2997
2	28900.79	28900.79	1

Figura 5.13: Número de elementos por clusters

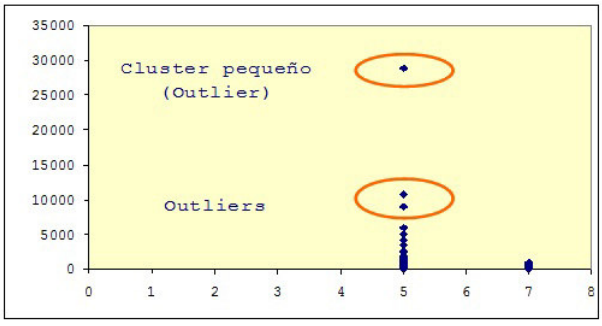

Figura 5.14: Outliers identificados con el método propuesto

Capítulo 6

Conclusiones

- El método propuesto, demuestra ser efectivo para detectar operaciones efectuadas con tarjetas de crédito con un comportamiento anómalo. Para efectuar ésta detección, en la primera fase del método se ejecuta iterativamente el algoritmo K-Means, con la finalidad de ubicar los mejores centros en cada cluster o agrupación de datos. Una vez estabilizados los centros, se procede con la búsqueda de los elementos fuera de serie también conocidos como Outliers, considerando que los valores de K y el porcentaje de Outliers a detectar deben ser determinados previamente. Cabe mencionar que este método puede ser utilizado para detectar comportamientos fuera de serie en otros ámbitos operativos.

- La ubicación de los puntos más alejados, en base a la distancia promedio de cada cluster, permite uniformizar mejor los elementos de cada cluster así como detectar, con mejor certeza, los posibles Outliers favoreciendo así la formación de los clusers basados en centros. Es posible utilizar otros tipos de medida de la similaridad tales como Minwoski o Malahanobis.

- Del caso expuesto se concluye que, con la aplicación del método propuesto se ubicaron 3 posibles casos de fraude, en el dataset real tomado como muestra, los cuales se manifiestan con los importes de las operaciones respectivas: $S/.$ 10787.84, $S/.$ 9000.00 y $S/.$ 28900.79 , identificando claramente a los elementos fuera de los patrones comúnmente establecidos.

- Como trabajos futuros, pueden incluirse métodos de optimización basados en Algoritmos Genéticos o GRASP, por ejemplo, para optimizar en cuanto a efec-

tividad la fase de búsqueda de los mejores centros. En la fase de búsqueda de Outliers, se podría incluir la generación de nuevos grupos, de tal manera que los Outliers no se ubiquen de uno en uno, sino por grupos.

Bibliografía

Alonso, Gonzáles: «Aprendizaje inductivo no basado en el error. Métodos no supervisados: Agrupamiento». *Departamento de Informática.Universidad de Valladolid. España*, 2008.

Bandyopadhyay, S. y Maulik, U.: «An evolutionary technique based on K-Means algorithm for optimal clustering in R^n». *Information Sciences*, 2002, **146(1–4)**, pp. 221–237.

Bezdek, J.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

Bhatla, T.Paul: «Understanding Credit Card Frauds». *Cards Business Review*, 2003.

Bradley, P. y Fayyad, U.: «Refining initial points for K-Means clustering». En: *Proc. 15th International Conf. on Machine Learning*, pp. 91–99. Morgan Kaufmann, San Francisco, CA, 1998.

Brucker, P.: «On the Complexity of Clustering Problems». *Lecture Notes in Economics and Mathematical Systems*, 1978, **157**, pp. 45–54.

Cano, J.; Cordón, O.; Herrera, F. y Sánchez, L.: «Greedy Randomized Adaptive Search Procedure Applied to the Clustering Problem as an Initialization Process Using K-Means as a Local Search Procedure». *International Journal of Intelligent and Fuzzy Systems*, 2002, **12**, pp. 235–242.

Chandola, Varun: «Outlier Detection - A Survey», 2007.

Chavez, E.; Navarro, G.; Baeza-Yates, R. y Marroquín, J.: «Searching in Metric Spaces». *ACM Compunting Surveys*, 2001, **33(3)**, pp. 273–321.

Dempster, A.; Laird, N. y Rubin, D: «Maximum-likelihood from imcomplete data via the em algorithm». *Journal of the Royal Statistical Society B*, 1977, **39**, pp. 1–39.

Dunn, J.: «A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters». *J. Cybernet*, 1974, **3**, pp. 32–57.

Elkan, C.: «Using the Triangle Inequality to Accelerate K-Means». En: *Proceedings of the Twentieth International Conference on Machine Learning (ICML '03)*, pp. 147–153, 2003.

Fayyad, U.; Piatetsky-Shapiro, G. y S., Padhraic: «From Data Mining to Knowledge Discovery in Databases». *American Association for Artificial Inteligence*, 1996, pp. 37–54.

Feo, T. y Resende, M.: «Greedy Randomized Adaptative Search Procedure». *Journal of Global Optimization*, 1995, **6**, pp. 109–133.

Forgy, E.: «Cluster analysis of multivariate data: Efficiency vs. Interpretability of classifications». *Biometrics*, 1965, **21**.

Garcia, Noelia: «Uso Fraudulento de Tarjetas Bancarias, http://delitosinformaticos.com/estafas/estafa$_t arjeta2.shtml \gg$, 2003.

Garey, M. y Johnson, D.: *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

Goldberg, D.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley Longman, Inc., 1989.

Hansen, P. y Mladenovic, N.: «J-Means: A new local search heuristic for minimun sum-of-squares clustering». *Pattern Recognition*, 2001, **34(2)**, pp. 405–413.

Hernandez, Edna: «Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto». *Tesis. Agosto 2006. México*, 2006.

Jain, A.; Murty y P., M. Flynn: «Data Clustering: a Review». *ACM Computer Surveys*, 1999, **31(3)**, pp. 264–323.

Kaufman, L. y Rousseeuw, P.: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York, NY, 1990.

KEARNS, M.; Y., MANSOUR y NG, A: «An Information-Theoretic Analysis of Hard and Soft Assignment Methods for Clustering». pp. 282–293. Morgan Kaufmann, 1997.

LIKAS, A.; VLASSIS, N. y VERBEEK, J.: «The Global K-means Clustering Algorithm». *Pattern Recognition*, 2003, **36**, pp. 451–461.

LUNG, CHUNG-HORNG; ZAMAN, MARZIA y NANDI, AMIT: «Applications of clustering techniques to software partitioning, recovery and restructuring». *J. Syst. Softw.*, 2004, **73(2)**, pp. 227–244. ISSN 0164-1212. doi: http://dx.doi.org/10.1016/S0164-1212(03)00234-6.

MARIN, JUAN MIGUEL: «Análisis de Cluster y Multidimensional Scaling. Análisis Multivariante, http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema5am.pdf». *Material docente Universidad Carlos III Madrid, diplomatura en Estadística.España*, 2006, **5**.

MAULIK, U. y BANDYOPADHYAY, S.: «Genetic algorithm-based clustering technique». *Pattern Recognition*, 2000, **33**, pp. 1455–1465.

MCQUEEN, J.: «Some methods for classification and analysis of multivariate observations». *In Preccedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

MERZ, P.: «Analysis of gene expression profiles: an application of memetic algorithms to the minimum sum-of-squares clustering problem». *BioSystems*, 2003, **72(1-2)**, pp. 99–109.

MOSCATO, P.: «On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms». *Caltech Concurrent Computation Program 826*, California Institute of Technology, Pasadena, California, USA, 1989.

MURTHY, C. y CHOWDHURY, N.: «In search of optimal clusters using genetic algorithms». *Pattern Recognition Letters*, 1996, **17**, pp. 825–832.

NEWMAN, GRAEME R.: «The Problem of Check and Card Fraud. Check and Card Fraud Guide No. 21 - 2003, http://www.popcenter.org/problems/credit$_c ard_f raud$/1 ≫ .*Material docente Universidad Carlos III Madrid, diplomatura en Estadística.España*, 2003, **21**.

ONTSIII: «Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información http://www.ontsi.red.es/empresas/indicator/id/520/medios-pago-utilizados-internet.html». *Estudio sobre Comercio Electrónico B2C 2007 (Octubre 2008)*, 2007.

Pacheco, J. y Valencia, O.: «Design of hybrids for the minimim sum-of-squares clustering problem». *Computational Statistics & Data Analysis*, 2003, **43(2)**, pp. 235–248.

Patel, Sukesh; Chu, William y Baxter, Rich: «A measure for composite module cohesion». En: *ICSE '92: Proceedings of the 14th international conference on Software engineering*, pp. 38–48. ACM Press, New York, NY, USA. ISBN 0-89791-504-6, 1992. doi: http://doi.acm.org/10.1145/143062.143086.

Patroni, Ursula: «EL PAGO ELECTRÓNICO, http://www.teleley.com/articulos/art$_p ago_e lectronico.pdf \gg, 2003 nalpoint$

Peña, J.; J., Lozano y Larrañaga, P: «An empirical comparison of four initialization methods for the K-Means algorithm». *Pattern Recognitioin Letters*, 1999, **20**, pp. 1027–1040.

Saeed, M.; Maqbool, O.; Babri, H.A.; Hassan, S.Z. y Sarwar, S.M.: «Software Clustering Techniques and the Use of Combined Algorithm». volumen 00, p. 301. IEEE Computer Society, Los Alamitos, CA, USA. ISBN 0-7695-1902-4, 2003. doi: http://doi.ieeecomputersociety.org/10.1109/CSMR.2003.1192438t

Wiggerts, T. A.: «Using Clustering Algorithms in Legacy Systems Remodularization». En: *WCRE '97: Proceedings of the Fourth Working Conference on Reverse Engineering (WCRE '97)*, p. 33. IEEE Computer Society, Washington, DC, USA. ISBN 0-8186-8162-4, 1997.